Fu Lee Wang
Zhiguo Gong
Xiangfeng Luo
Jingsheng Lei (Eds.)

LNCS 6318

Web Information Systems and Mining

International Conference, WISM 2010
Sanya, China, October 2010
Proceedings

Springer

Fu Lee Wang Zhiguo Gong Xiangfeng Luo
Jingsheng Lei (Eds.)

Web Information Systems and Mining

International Conference, WISM 2010
Sanya, China, October 23-24, 2010
Proceedings

Springer

Volume Editors

Fu Lee Wang
Caritas Francis Hsu College, Department of Business Administration
18 Chui Ling Road, Tseung Kwan O, Hong Kong, China
E-mail: pwang@cihe.edu.hk

Zhiguo Gong
University of Macau, Department of Computer and Information Science
Av. Padre Tomás Pereira, Taipa, Macau, SAR, China
E-mail: fstzgg@umac.mo

Xiangfeng Luo
Shanghai University, School of Computer
99 Shangda Road, Shanghai 200444, China
E-mail: luoxf@shu.edu.cn

Jingsheng Lei
Nanjing University of Posts and Telecommunications, School of Computer
Nanjing 210003, China
E-mail: leijs@njupt.edu.cn

Library of Congress Control Number: 2010936446

CR Subject Classification (1998): H.4, H.3, H.5, H.2, C.2.4, I.2.6

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

ISSN	0302-9743
ISBN-10	3-642-16514-1 Springer Berlin Heidelberg New York
ISBN-13	978-3-642-16514-6 Springer Berlin Heidelberg New York

springer.com

Printed in Germany

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

Preface

The 2010 International Conference on Web Information Systems and Mining (WISM 2010) was held October 23–24, 2010 in Sanya, China. WISM 2010 received 603 submissions from 20 countries and regions. After rigorous reviews, 54 high-quality papers were selected for publication in the WISM 2010 proceedings. The acceptance rate was 9%.

The aim of WISM 2010 was to bring together researchers working in many different areas of Web information systems and Web mining to foster the exchange of new ideas and promote international collaboration. In addition to the large number of submitted papers and invited sessions, there were several internationally well-known keynote speakers.

On behalf of the Organizing Committee, we thank Hainan Province Institute of Computer and Qiongzhou University for its sponsorship and logistics support. We also thank the members of the Organizing Committee and the Program Committee for their hard work. We are very grateful to the keynote speakers, invited session organizers, session chairs, reviewers, and student helpers. Last but not least, we thank all the authors and participants for their great contributions that made this conference possible.

October 2010

Fu Lee Wang
Gong Zhiguo
Xiangfeng Luo
Jingsheng Lei

Organization

Organizing Committee

General Co-chairs

Qing Li	City University of Hong Kong, China
Mingrui Chen	Hainan University, China

Program Committee

Co-chairs

Gong Zhiguo	University of Macau, Macau
Xiangfeng Luo	Shanghai University, China

Local Arrangement

Chairs

Zhuang Li	Qiongzhou University

Proceedings

Co-chair

Fu Lee Wang	Caritas Francis Hsu College, China
Jingsheng Lei	Nanjing University of Posts and Telecommunications, China

Publicity

Chair

Lanzhou Wang	China Jiliang University, China

Sponsorship

Chair

Zhiyu Zhou	Zhejiang Sci-Tech University, China

Program Committee

Ladjel Bellatreche	ENSMA - Poitiers University, France
Sourav Bhowmick	Nanyang Technological University, Singapore
Stephane Bressan	National University of Singapore, Singapore
Erik Buchmann	University Karlsruhe, Germany
Jinli Cao	La Trobe University, Australia
Jian Cao	Shanghai Jiao Tong University, China
Badrish Chandramouli	Microsoft Research, USA
Akmal Chaudhri	City University of London, UK
Qiming Chen	Hewlett-Packard Laboratories, USA
Lei Chen	Hong Kong University of Science and Technology, China
Jinjun Chen	Swinburne University of Technology, Australia
Hong Cheng	The Chinese University of Hong Kong, China
Reynold Cheng	Hong Kong Polytechnic University, China
Bin Cui	Peking University, China
Alfredo Cuzzocrea	University of Calabria, Italy
Wanchun Dou	Nanjing University, China
Xiaoyong Du	Renmin University of China, China
Ling Feng	Tsinghua University, China
Cheng Fu	Nanyang Technological University, Singapore
Gabriel Fung	The University of Queensland, Australia
Byron Gao	University of Wisconsin, USA
Yunjun Gao	Zhejiang University, China
Bin Gao	Microsoft Research, China
Anandha Gopalan	Imperial College, UK
Stephane Grumbach	INRIA, France
Ming Hua	Simon Fraser University, Canada
Ela Hunt	University of Strathclyde, Glasgow
Renato Iannella	National ICT, Australia
Yan Jia	National University of Defense Technology, China
Yu-Kwong Ricky	Colorado State University, USA
Yoon Joon Lee	KAIST, South Korea
Carson Leung	The University of Manitoba, Canada
Lily Li	CSIRO, Australia
Tao Li	Florida International University, USA
Wenxin Liang	Dalian University of Technology, China
Chao Liu	Microsoft, USA
Qing Liu	CSIRO, Australia

Jie Liu	Chinese Academy of Sciences, China
JianXun Liu	Hunan University of Science and Technology, China
Peng Liu	PLA University of Science and Technology, China
Jiaheng Lu	University of California, Irvine, USA
Weiyi Meng	Binghamton University, USA
Miyuki Nakano	University of Tokyo, Japan
Wilfred Ng	Hong Kong University of Science and Technology, China
Junfeng Pan	Google, USA
Zhiyong Peng	Wuhan University, China
Xuan-Hieu Phan	University of New South Wales (UNSW), Australia
Tieyun Qian	Wuhan University, China
Kaijun Ren	National University of Defense Technology, China
Dou Shen	Microsoft, USA
Peter Stanchev	Kettering University, USA
Xiaoping Su	Chinese Academy of Sciences, China
Jie Tang	Tsinghua University, China
Zhaohui Tang	Microsoft, USA
Yicheng Tu	University of South Florida, USA
Junhu Wang	Griffith University, Australia
Hua Wang	University of Southern Queensland, Australia
Guoren Wang	Northeastern University, USA
Lizhe Wang	Research Center Karlsruhe, Germany
Jianshu Weng	Singapore Management University, Singapore
Raymond Wong	Hong Kong University of Science and Technology, China
Jemma Wu	CSIRO, Australia
Jitian Xiao	Edith Cowan University, Australia
Junyi Xie	Oracle Corp., USA
Wei Xiong	National University of Defense Technology, China
Hui Xiong	Rutgers University, USA
Jun Yan	University of Wollongon, Australia
Xiaochun Yang	Northeastern University, China
Jian Yang	Macquarie University, Australia
Jian Yin	Sun Yat-Sen University, China
Qing Zhang	CSIRO, Australia
Shichao Zhang	University of Technology, Australia
Yanchang Zhao	University of Technology, Australia
Sheng Zhong	State University of New York at Buffalo, USA
Aoying Zhou	East China Normal University, China
Xingquan Zhu	Florida Atlantic University, USA

Table of Contents

Applications of Web Information Systems

Applications of Web Mining

Distributed Systems

E-Government and E-Commerce

Geographic Information Systems

Information Security

Intelligent Networked Systems

Management Information Systems

Mobile Computing

Web Content Mining

Web Information Classification

Web Information Retrieval

Web Services and E-Learning

XML and Semi-structured Data

An Improving Multi-Objective Particle Swarm Optimization

JiShan Fan

Huai Institute of Technology, School of Electronic Engineering
LianYunGang China 222005
fjsszw2005@126.com

Abstract. In the past few years, a number of researchers have successfully extended particle swarm optimization to multiple objectives. However, it still is an important issue to obtain a well-converged and well-distributed set of Pareto-optimal solutions. In this paper, we propose a fuzzy particle swarm optimization algorithm based on fuzzy clustering method and fuzzy strategy and archive update. The particles are evaluated and the dominated solutions are stored into different cluster in the generation, while dominated solutions are pruned. The non-dominated solutions are selected by fuzzy strategy, and the non-dominated solutions are added to the archive. It is observed that the proposed fuzzy particle swarm optimization algorithm is a competitive method in the terms of convergence near to the Pareto-optimal front, diversity of solutions.

Keywords: particle swarm optimization; fuzzy clustering method; fuzzy strategy.

1 Introduction

The particle swarm optimization (PSO) was proposed by Kennedy and Eberhart in 1995[1], its inception has gained rapid popularity as a method to facilitate single objective optimization. Like genetic algorithms, PSO was inspired by nature, but instead of evolution it was the flocking and swarm behavior of birds and insects that motivated its development.

Due to the success of PSO in single objective optimization, in recent years, lots of researchers have made to extend PSO to the domain of multi-objective problems, so many PSO (MOPSO) algorithms were published [3, 4, 5, 6]. Although most of these MOPSO algorithms were generated, there were lots of shortages. Especially, they obtained a well-converged and well-distributed set of Pareto-optimal solutions in multi-objective evolutionary optimization.

In PSO itself the swarm population is fixed in size, and its members cannot be replaced, only adjusted by their global best individual and personal best individual [2].However, in order to facilitate an multi-objective optimization approach to PSO a set of non-dominated solutions must be replace the single global best individual .So, we select the non-dominated solutions by the clustering method and fuzzy strategy. The algorithm is observed in terms of convergence near to the Pareto-optimal front and diversity of solutions.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 1–6, 2010.

2 Fuzzy Clustering Method

In addition to the convergence to the Pareto-optimal front, one of the important aspects of multi-objective optimization is to find and maintain a widely distributed set of solutions. Since the Pareto-optimal front can be convex, non-convex, disconnected, or piece-wise continuous hyper-surfaces, so there are difficulties in maintaining diversity among population members. In order to well choose non-dominated solutions, we select clustering method to estimate dominated individuals.

Step1. Confirm the fuzzy connection of particles by fitness value;
Step2. Compute the distance of particles by the flowing expression:

$$R(x_i,x_j)=\left|\frac{\min_{1\le k\le N}(dis\tan ce(x_i,x_k))-dis\tan ce(x_i,x_j)}{\min_{1\le k\le N}(dis\tan ce(x_i,x_k))-\min_{1\le k\le N}(dis\tan ce(x_j,x_k))}\right| \quad (1)$$

In above expression (1), the distance is between swarm member x_i and another cluster member x_k, N is the dimension of objective function, it is calculated as:

$$distacne(x_i,x_k)=\left|\sum_{i=1}^{N}\left(\frac{f_i(x_i)-f(x_k)}{\max(f_i(X))-\min(f_i(X))}\right)\right| \quad (2)$$

By dividing each of the particle objective distances by its range in the cluster, the distance space is normalized, so this can mitigate any objective scaling differences.

Step3. Select a parameter $\lambda[0,1]$, if $R(x_i,x_j)\ge\lambda$, then x_i and x_j are in same cluster, else go to Step2.

Passing the above calculate process, the particles of any objective function are classified into different clusters.

3 Fuzzy Strategy

In order to obtain a well-converged and well-distributed set of Pareto-optimal solutions, we propose a fuzzy Strategy to select non-dominated particle solutions.

Step1. Order the particle individuals by their fitness value in the different cluster;
Step2. Fuzzy mapping the values of multi-objective by the flowing formula:

$$\xi_F(x_i)=\begin{cases}1-\lambda*\dfrac{D}{P} & x\le u\\ 0 \quad x>u\end{cases} \quad (3)$$

In the formula, is the value of objective function, D is the number of multi-objective functions, P is the ordering number of particle individuals, u is the threshold, is a parameter, its value is between 0 and 1.
Step3. Compute the velocity and position of each particle in a set of M particles.

4 Fuzzy Multi-Objective Particle Swarm Optimization (FMOPSO)

In multi-objective particle swarm optimization, we propose the clustering method and fuzzy strategy in the MOPSO [7, 8,9].

The multi-objective particle swarm optimization begins with a set of uniformly distributed random initial particles defined in the search space S.

A set of M particles are considered as a population P. Each particle i has a position p_i and a velocity v_i. The position p_i was defined by $x^i = (x_1^i, x_2^i, ..., x_n^i)$ and the velocity v_i was defined by $v^i = (v_1^i, v_2^i, ..., v_n^i)$ in the search space S.

Beside the population P, another set (called Cluster) C can be defined in order to select non-dominated solutions, in each cluster, there are many dominated particles with fuzzy connection, another set (called Archive) A can be defined in order to store the obtained non-dominated solutions. Due to the presence of A, better individuals are preserved during generations.

The particles are evaluated and the non-dominated solutions are select from C. In the next step, the particles are moved to a new position in the space S. The position and the velocity of each particle i is updated by the flowing formulation:

$$v_{j,t+1}^i = wv_{j,t}^i + \xi_F(x_{j,t}^i) * (c1(p_{j,t}^i - x_{j,t}^i) + c2(p_{j,t}^{i,g} - x_{j,t}^i)) \tag{4}$$

$$x_{j,t+1}^i = x_{j,t}^i + v_{j,t+1}^i. \tag{5}$$

Where j=1, 2,…,N, i=1,2,…,M, c1 and c2 are two positive constants, w is the inertia weight which is employed to control the impact of the previous history of velocities. $p_t^{i,g}$ is the position of the global best particle in the population. p_t^i is the best position that particle i can find so far and keeps the non-dominated position of the particle by comparing the new position x_{t+1}^i in the objective space.

5 Simulation and Results

5.1 Test Problems

In order to compare the scalability behavior of the IMODE algorithm, we choose five well known test functions, namelyZDT1, ZDT2, ZDT3, ZDT4 and ZDT6 [10].

5.2 Experiments and Results

To make FMOPSO algorithm powerful on global optimization in the early evolution process and to get good convergence performance in the later period, we set parameters of algorithm.

Parameter Setting: The FMOPSO method is used with 100 particles and an internal archive size of 100 and an internal cluster size of 100 and is run 100 generations. We select values for threshold u and parameter λ as 0.01 and 0.1. At the same time, we select standard values for turbulence factor and inertia weights as 0.1 and 0.4.

Evaluations: The optimal solutions are showed from Figure 1 to Figure5.

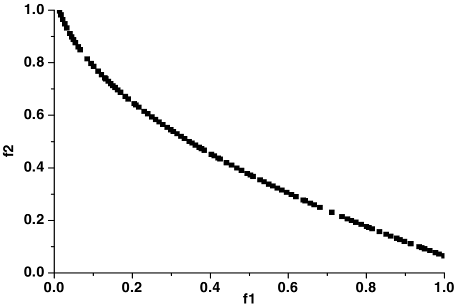


Fig. 1. The Optimal results of ZDT1

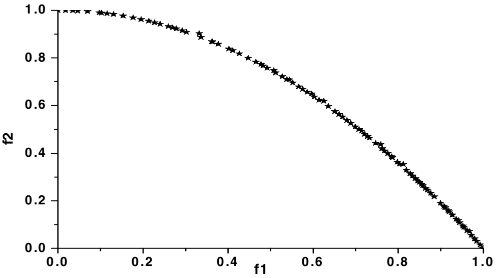


Fig. 2. The Optimal results of ZDT2

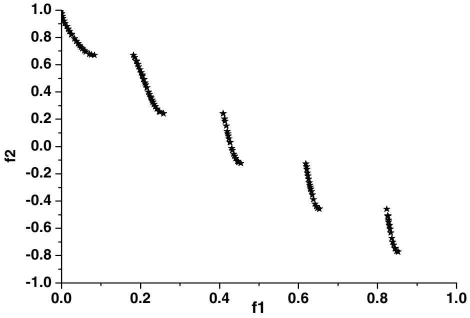


Fig. 3. The Optimal solutions of ZDT3

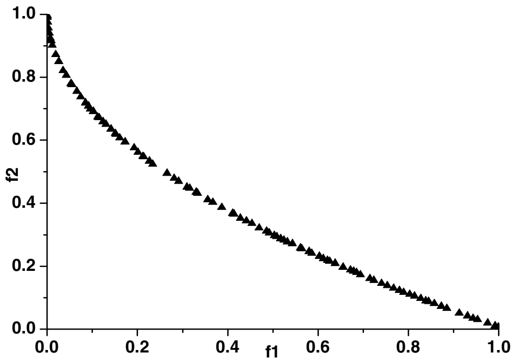


Fig. 4. The Optimal Solutions of ZDT4

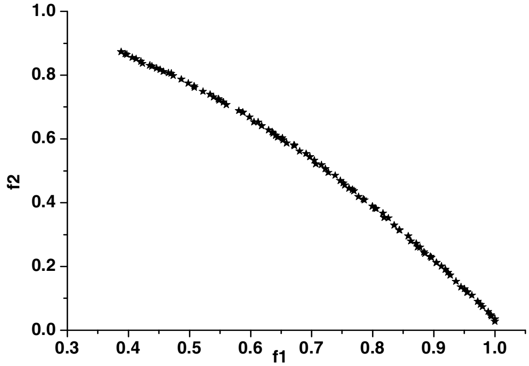


Fig. 5. The Optimal Solutions of ZDT6

In the fig. 1-5, the results are optimized by using MOPSO algorithms respectively in order to explain FMOPSO algorithms possess of better performance in optimization problems than algorithms.

5.3 Conclusion and Discussion

ZDT1, ZDT2 and ZDT3 are high-dimensionality of Multi-objective problems. Many MOEAs have achieved very good results on these problems in convergence to the true Pareto front and uniform spread of solutions along the front [11, 12, 13, 14, 15]. The results for three problems show that MOPSO achieves better results, the challenge for FMOPSO in the ZDT4 problem; because it has many local Pareto fronts (2^{19}) that tend to mislead the optimization algorithm. From the results of FMOPSO, the algorithm shows better performance in the hard optimization problem.

The test problem ZDT6 has two enormous difficulties. They are thin density of solutions towards the true Pareto front and non-uniform spread of solutions along the Pareto front. From the Figure 5, these algorithms show the similar results of the ZDT6.

With the results of the simulation excrement for five well known optimization problems, the FMOPSO algorithm shows better performance.

References

[1] Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of the Fourth IEEE International Conference on Neural Networks, Perth, Australia, pp. 1942–1948 (1995)

[2] Fieldsend, J.E.: Multi-objective particle swarm optimization methods (2004)

[3] Coello, C.A.C., Lechunga, M.S.: MOPSO: A Proposal for Multiple Objective Particle Swarm Optimizations. In: Proceedings of the 2002 Congress on Evolutionary Computation, part of 2002 IEEE World Congress on Computational Intelligence, Hawaii, May 12-17, pp. 1051–1056 (2002)

[4] Fieldsend, J.E., Singth, S.: A Multi-Objective Algorithm based upon Particle Swarm Optimisation, an Efficient Data Structure and Turbulence. In: Proceedings of UK Workshop on Computational Intelligence, Birmingham,UK, September 2-4, pp. 37–44 (2002)

[5] Hu, X., Eberthart, R.: Multiobjective Optimization Using Dynamic Neiborhood Particle Swarm Optimization. In: Proceedings of the 2002 Congress on Evolutionary Computation, part of the 2002 IEEE world Congress on Computational Intelligence, Hawii, May 12-17 (2002)

[6] Parsopoulos, K.E., Vrahatis, M.N.: Particle Swarm Optimization Method in Multiobjective Problems. In: Proceedings of the 2002 ACM Symposium on Applied Computing, pp. 605–607 (2002)

[7] Branke, J., Kamper, A., Schmeck, H.: Distribution of evolutionary algorithms in heterogeneous networks. In: Deb, K., et al. (eds.) GECCO 2004. LNCS, vol. 3102, pp. 923–934. Springer, Heidelberg (2004)

[8] Mostaghim, S., Teich, J.: Covering pareto-optimal fronts by subswarms in multi-objective particle swarm optimization. In: IEEE Proceedings, World Congress on Computational Intelligence(CEC 2004), Portland, USA, pp. 1404–1411 (June 2004)

[9] Mehnen, J., Michelitsch, T., Schmitt, K., Kohlen, T.: pMOHypEA: Parallel evolutionary multiobjective optimization using hypergraphs. Interner Bericht des Sonderforschungsbereichs 531 Computational Intelligence CI–189/04, Universität Dortmund (Dezember 2004)

[10] Zitzler, E., Deb, K., Thiele, L.: Comparison of multi- objective evolutionary algorithms: Empirical results, Evolutionary Computation, pp. 173–195 (2000)

[11] Abbass, H.A., Sarker, R., Newton, C.: PDE: A pareto-frontier differential evolution approach for multi-objective optimization problems. In: Proceedings of the Congress on Evolutionary Computation 2001 (CEC 2001), vol. 2, pp. 971–978. IEEE Service Center, New Jersey (2001)

[12] Madavan, N.K.: Multiobjective optimization using a pareto differential evolution approach. In: Congress on Evolutionary Computation (CEC 2002), vol. 2, pp. 1145–1150. IEEE Service Center, New Jersey (2002)

[13] Zitzler, E., Deb, K., Thiele, L.: Comparison of multi- objective evolutionary algorithms: Empirical results, Evolutionary Computation, pp. 173–195 (2000)

[14] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and Elitist Multi-objective Genetic Algorithm: NSGA_II. IEEE Transaction on Evolutionary Computation 6(2), 182–197 (2002)

[15] Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro-machine and Human Science, pp. 39–43 (1995)

Study on RSA Based Identity Authentication Algorithm and Application in Wireless Controlling System of Aids to Navigation

Fu-guo Dong and Hui Fan

School of computer science and technology,
Shandong Institute of Business and Technology, Yantai, China
dongfuguo2005@126.com, fanlinw@263.net

Abstract. This paper studies the safe wireless communication of controlling instructions in system of aids to navigation. RSA based authentication algorithm is employed, and Euclid addition chains are used to compute the modular exponentiation of RSA, transforming the shortest addition chains of a large number e into those of three smaller numbers a, b and c, which meets e=a×b+c. Experiment results show that the Euclid addition chains need much less time and space to create. In the case of wireless communication with authentication, controlling instructions can be transmitted safely and accurately, and it is hard for malicious users to control the intelligent lights directly. This is of great significance for navigation trades.

Keywords: System of Aids to Navigation; RSA; Authentication Algorithm; Euclid Addition chains.

1 Introduction

In telemetering and telecontrol system of aids to navigation, controlling component of intelligent lights communicates with the controlling center mainly by GSM/GPRS wireless network. Users can send instructions through to the intelligent lights at any time to query and modify the status, including electricity, voltage, and other parameters. Intelligent lights can communicate with controlling center freely, this can not only reduce the times and costs for inspection and maintenance, and more important can detect and solve problems in times so that many traffic accidents are avoided. Intelligent lights can be managed unattended, and people are released from the island. But, because of openness of wireless network, malicious users can control the intelligent lights without through the correct controlling center. Once incorrectly controlled, the lights will not be able to provide navigation information accurately, which is dangerous to the shipping traffic. It is greatly necessary to ensure that the wireless communication network between intelligent lights and controlling center is secret and accurate, especially the controlling instruction communications. The intelligent lights must be able to know whether an instruction is from the correct controlling center or not. It is greatly necessary for intelligent lights to validate the identity of controlling center when receiving every controlling instructions.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 7–13, 2010.

According to the asymmetry of private key and public key, RSA can be used to design identity authentication and digital signature systems to transmit information safely in public channels. To ensure complete security, controlling component of intelligent lights should know whether an instruction is from the correct controlling center, and the controlling center should also know whether any information is from the correct intelligent lights. That is, two-way identity authentication should be employed in system of aids to navigation. But on the other hand, considering the fluency of the whole system, instantaneity of query instructions, accuracy of controlling instructions, and working highlights of the system, and many other factors, we only design one-way identity authentication. The server of controlling center sends the query instructions directly, and adds encrypted identity information to the instruction datagram before sending controlling instruction. When receiving query instructions from controlling center, intelligent lights will return its status immediately. When receiving controlling instruction, intelligent lights will validate it, if it is from the correct controlling center, then intelligent lights will modify its status according to the instruction demand and return new status, otherwise, the instruction will be dropped directly. In this way, instantaneity and accuracy are ensured at the same time. In this paper, we give a scheme for navigation intelligent lights to validate the identity of the controlling center, using RSA algorithm, especially in the case of controlling instruction. Euclid addition chains are used to compute modular exponentiation of RSA. A lot of experiment results show that the scheme we give can get satisfactory performance[1-3].

2 Datagram Format

Controlling components of intelligent lights must be able to know whether an instruction is a query one or a controlling one, and more important to know whether a controlling instruction is from the correct controlling center or not. All these must be tagged in the format of datagram. So the format of datagram is designed as follows: GPRS Header: Instruction Type: Encrypted Identifier: Content.

In application, the value 11110000 in “Instruction Type” standards for query instruction, and the value 00001111 standards for controlling instruction, any other value means invalid instructions and will be dropped directly. The “encrypted Identifier” field contains the encrypted identifier of controlling center. The server of controlling center sends the query instructions directly, and adds encrypted identity information to the instruction datagram before sending controlling instruction. After receiving instructions, controlling component of intelligent lights will unpack the datagram according to GPRS protocol, and then judge the instruction type. If it is a query instruction, the “Encrypted Identifier” field will be ignored, and the current status will be returned immediately. If it is a controlling instruction, the controlling component of intelligent lights will validate with the “Encrypted Identifier”. If successfully authenticated, intelligent lights will then modify the status of itself, and return new status. Otherwise, if the instruction if neither a query one nor a controlling one, it will be dropped directly. The “Content” field contains the content of instructions.

3 RSA-Based Authentication Algorithm and Its Application

In telemetering and telecontrol system of aids to navigation, only the controlling components of navigation intelligent lights need to authenticate the identity of controlling center. So, only the public key and private key as well as the identifier of the controlling center need to be designed. Public key and identifier of controlling center are installed on controlling component of intelligent lights; private key and identifier of controlling center are installed on the server of controlling center. When sending status query instruction, the server of controlling center will randomly fill the "Encrypted Identifier" field of instruction datagram; when sending controlling instruction, controlling center at first encrypt the identifier using private key, and then fill the "Encrypted Identifier" field with it. Because of the confidentiality of private key and the corresponding relationship of public key and private key, it is hard for malicious users to fabricate private key, which avoids the direct control of intelligent lights, ensures the steady working of intelligent lights. Flow chart of controlling center sending instructions is as Fig 1.

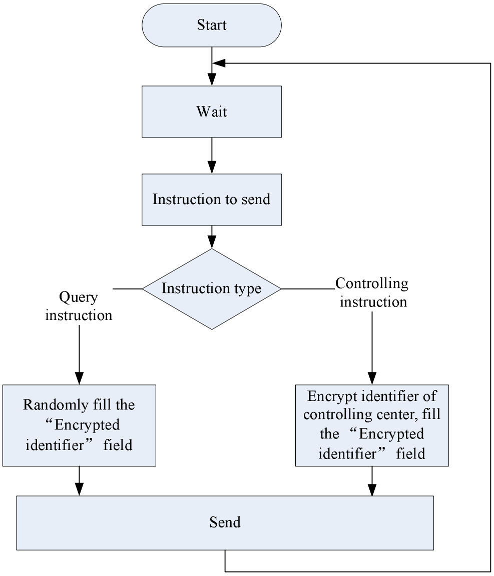

Start
Wait
Instruction to send
Instruction type
Query instruction
Controlling instruction
Randomly fill the "Encrypted identifier" field
Encrypt identifier of controlling center, fill the "Encrypted identifier" field
Send

Fig. 1. Illustration of controlling center sending instruction

Fig 2 illustrates the flow chart of controlling component of intelligent lights receiving instructions.

3.1 Choice of RSA Key

After created, public key can be let known to everyone, while private key should be strictly kept secret. In application, private key is generally designed very large, while

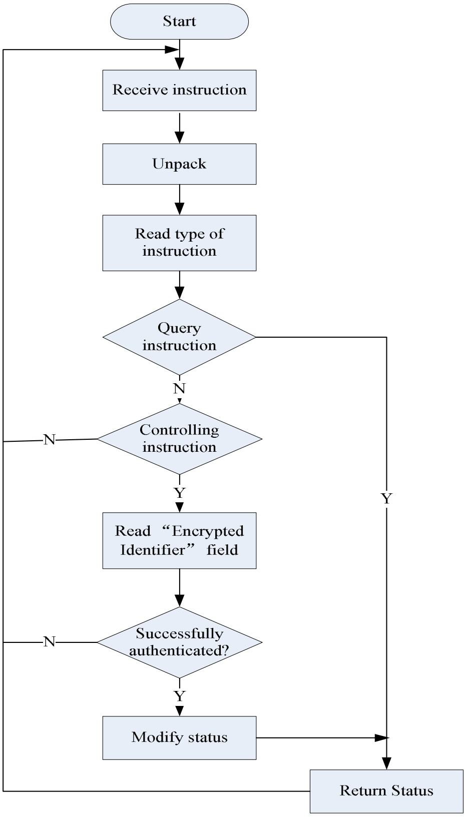


Fig. 2. Illustration of program after receiving instruction

public key is smaller and simpler, so that fast authentication is possible. And in system of aids to navigation, intelligent lights are installed on island far from land, the condition of power supply is very poor, and it is not a good idea to put more computation on intelligent lights. So public key should be chosen smaller. On the other hand, the server of controlling center has strong power to do computations and power supply is enough. So private key is generally very large and complex[4-8].

3.2 Choice of Modular Exponentiation Algorithm

According to Euclid algorithm, for any integer e, there exist three integers a, b, c satisfy

$$e = a \times b + c \tag{1}$$

Then we have

$$m^e = m^{a \times b + c} = (m^a)^b \cdot m^c = (m^b)^a \cdot m^c \tag{2}$$

Algorithm analysis and a lot of experiment results show that for any given integer e, there are three integers a ,b ,c (c may be 0), which meet

$$T(e)=\begin{cases} T(a)+T(b)+T(c)+1, c\neq 0 \\ T(a)+T(b) \qquad\qquad ,c=0 \end{cases} \tag{3}$$

Where $T(x)$ denotes the multiplication times of shortest addition chains of x. If c is a node of the shortest addition chains of a or b, then m^c can be gotten by the way during the course of computing m^a or m^b, we have

$$T(e)=\begin{cases} T(a)+T(b)+1, c\neq 0 \\ T(a)+T(b) \qquad ,c=0 \end{cases} \tag{4}$$

From above we can see that transforming the shortest addition chains of e into those of three little integers a ,b ,c is more feasible and significant. If c equals 0 or is a child chain of a or b, this algorithm will achieve highest efficiency. Suppose the shortest addition chains of a is $a_0, a_1, \cdots, c, \cdots a_m$, that of b is $b_0, b_1, \cdots b_n$, and c is a child node of the shortest addition chains of a. Then m^c can be gotten during computing m^a, so many multiplication can be reduced.

In system of aids to navigation, controlling component of intelligent lights and the server of controlling center use Euclid based addition chains algorithm for modular exponentiation computation, and do not store the real private key and public key. If malicious users want to compute the real private key, they have to do a great lot of attempts, which can not finished in limited time, even if three addition chains are let out. Then the security of private key is ensured, and the security and reliability of wireless communication are guaranteed[9].

3.3 Choice of Identifier of Controlling Center

To ensure the security and reliability of wireless communication of controlling instructions, it is necessary to generate identifier of controlling center and to install it on controlling component of intelligent lights and the server of controlling center. When sending controlling instructions, the server of controlling center should encrypt the identifier using private key and fill the “Encrypted Identifier” field of datagram, which is transmitted through wireless network of GPRS. When receiving controlling instructions, the controlling component of intelligent lights will read and decrypt the “Encrypted Identifier” field to get the received identifier and compare it with the installed one. All these need computation time and wireless bandwidth. Then the identifier should be carefully chosen so that the “Encrypted Identifier” field can be limited in suitable length. In this system, we choose a 50-digit length random number as the identifier of controlling center, and ensure that the “Encrypted Identifier” field is less than 30-digit length.

4 Simulation and Analysis of Communication Performance

In experiments, we use Visual C# 2008 to simulate the wireless communication with and without identity authentication. In the case of communication of controlling instructions with authentication, Euclid based addition chains are used to compute modular exponentiation of RSA.

Experiment results show that the length of datagram become larger by 100 bits after employing authentication. For GPRS communication at the speed of 54.4kpbs, this will almost not influence the performance. The time delay is mainly caused by the computation of modular exponentiation, while the time needs to judge the type of instruction and to compare the received identifier with the already installed one can be ignored. Table 1 illustrates the experiment results, for 1000 times of experiments of 400 intelligent lights and a controlling center. In this table, return time mainly consists of transmitting time and that of computation of modular exponentiation, as well as that of detecting and modifying the status of intelligent lights.

Table 1. Comparison of communication efficiency

Communication Mode	Instruction Type	Return Time(s)	Accuracy
Communication with authentication	query	0.137	100%
	controlling	3.103	100%
Communication without authentication	query	0.120	100%
	controlling	1.003	100%

As the table shows, the return time becomes slightly longer in the case of communication with authentication. In both cases the accuracy is guaranteed. But after employing identity authentication, controlling instructions can be transmitted safely and accurately, and it is hard for malicious users to control the intelligent lights. This is of great significance for navigation trades.

Acknowledgment

This work is supported by National Natural Science Foundation (60673153), National Natural Science Foundation (60773053), National Natural Science Foundation (60803048), Shandong Province Natural Science Foundation (Y2007A28), Shandong Outstanding Young Scientist Research Award Fund (2008BS01026), and Science and technology projects of Shandong Province Education Department No. J10LG21.

References

1. Bergeron, F., Berstel, J., Brlek, S., et al.: Addition Chains Using Continued Fractions. J. Algorithms 10, 403–412 (1989)
2. Rivest, R.L., Shamir, A., Adleman, L.M.: A Method for Obtaining Digital Signatures and Public Key Cryptosystems. Communications of the ACM 21, 120–126 (1978)

3. Koblitz, N.: A Course in Number Theory and Cryptography. Springer, New York (1987)
4. Rivest, R.L., Shamir, A., Adleman, L.M.: On Digital Signatures and Public Key Cryptosystems: [Technical Report]. MIT/LCS/ TR-212.MIT Laboratory for Computer Science (1979)
5. Knuth, D.E.: The Art of Computer Programming: Seminumerical Algorithms, 3rd edn., vol. 2. Addtion-Wesley, Reading (2003)
6. Bos, J., Coster, M.: Addition chain heuristics. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 400–407. Springer, Heidelberg (1990)
7. Kunihiro, N., Yamamoto, H.: New Methods for Generating Short Addition Chains. IEICE Trans. Fundamentals E83-A1, 60–67 (2000)
8. Montgomery, P.L.: Modular Multiplication without Trial Division. J. Mathematics of Computation 170, 519–521 (1985)
9. Fu-guo, D., Yu-rong, L.: A Novel Shortest Addition Chains Algorithm Based On Euclid Algorithm. In: Proceeding of WiCom (2008)

Invariant Subspaces for Some Compact Perturbations of Normal Operators*

Mingxue Liu

Faculty of Computer Science, Guangdong Polytechnic Normal University
Guangzhou 510665, People′s Republic of China
liumingxue9698@sina.com.cn

Abstract. The famous computer scientist J. von Neumann initiated the research of the invariant subspace theory and its applications. This paper show that compact perturbations of normal operators on an infinite dimensional Hilbert space H satisfying certain conditions have nontrivial closed invariant subspaces.

Keywords: Invariant subspace, normal operator, compact perturbation.

The invariant subspace theory has important applications in computer science and information science (see [1-13] and so on). As stated in [2], it was the famous computer scientist J. von Neumann who initiated the research of the invariant subspace theory for compact operators and its applications. To be more specific, J. von Neumann showed that compact operators on an infinite dimensional Hilbert space H have nontrivial closed invariant subspaces.

One of the most difficult question in the invariant subspace theory is the problem of existence of invariant subspaces for compact perturbations of self adjoint operators as well as normal operators (see [10] and so on). One of the long open question is the problem of existence of invariant subspaces for compact perturbations of normal operators (see [11] and so on) .

In 1996, A. Simonic showed in [13] that compact perturbations of self-adjoint operators on a real infinite dimensional Hilbert space H have nontrivial closed invariant subspaces. It is well known that any self-adjoint operator must be a normal operator, but the converse is not true.

In this paper, we extend the result of [13] to a wider class of not necessarily self-adjoint operators. To be more specific, we show that compact perturbations of normal operators on an infinite dimensional Hilbert space H satisfying certain conditions have nontrivial closed invariant subspaces.

Since the proof of our main result is based on ideas from [13], we shall freely use the result and notation appears in that paper. The key change appears in the lemmas whose proof need to use properties of self-adjoint operators. Moreover, the essential spectrum of a compact perturbation of a normal operator might be complex.

* The research was supported by the Natural Science Foundation of P. R. China (No.10771039).

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 14–18, 2010.

Let H be an infinite dimensional Hilbert space over the field F with $F = R$ (the field of real numbers) or $F = C$ (the field of complex numbers), and let $B(H)$ denote the algebra all bounded linear operators on H. For an operator $T \in B(H)$, T^* and $\|T\|_{\mathrm{e}}$ denote its Hilbert adjoint operator and the essential norm respectively; ReT stands for its real parts, that is, $\mathrm{Re}T = \frac{1}{2}(T + T^*)$. By(1.2.1) from [13], $\|\mathrm{Re}T\| \leq \|T\|_{\mathrm{e}} \leq \|T\|$.

If Π is a C^*-algebra, a linear functional $f : \Pi \longrightarrow F$ is positive if $f(a) \geq 0$ for all positive elements in Π. A state τ on Π is a positive linear functional on Π of norm 1. It is well known that if τ is a positive linear functional on the C^*-algebra Π, then τ is bounded and $\|\tau\| = \tau(1)$.

Lemma 1 ([13], Proposition 1.4.1). Let $\mathcal{A}$ be a convex subset of $B(H)$. Fixed a unit vector $x_0 \in H$, and choose a positive number $r \in (0, 1)$. Suppose that for every vector $y_0 \perp x_0$, and $\|y_0\| \leq 1$, there exists an operator $A \in\mathcal{A}$ satisfying the following strict inequality:

$$\mathrm{Re}\left\langle A(x_0 + \frac{r}{\sqrt{1-r^2}}y_0), x_0 - \frac{\sqrt{1-r^2}}{r}y_0 \right\rangle > \|\mathrm{Re}A\|_{\mathrm{e}}(1 - \|y_0\|^2).$$

Then $\mathcal{A}$ contains an operator A_0, with an eigenvector in the set

$$S = \{x \in H; \|x - x_0\| \leq r\},$$

and the corresponding eigenvalue μ satisfies the condition: $|\mathrm{Re}\mu| > \|\mathrm{Re}A_0\|_{\mathrm{e}}$.

From now on, suppose $A = N + K \in B(H)$ is a fixed operator without nontrivial closed invariant subspaces, where N is a normal operator, and K a compact operator. We denote by E the essential spectrum of A. Furthermore, we may assume that the norm $\|N\| \leq 1$. Consequently, we have $E \subset D := \{\mu \in F; |\mu| \leq 1\}$.

The algebra of all polynomials with the coefficients in F, equipped with the norm

$$\|p\|_\infty = \sup\{|p|(\mu)|; \mu \in E\}$$

is denoted by $\mathcal{P}(E)$.

Definition 1 ([13], Definition 2.3.1). Let $D \subset H$ be the set of all nonzero vector $x \in H$ for which there exists a nonzero vector $y \in H$ such that the following inequality

$$\mathrm{Re}\langle p(A)x, y\rangle \leq \|p\|_\infty \langle x, y\rangle$$

holds for every polynomial $p \in \mathcal{P}(E)$.

Lemma 2 ([13], Lemma 2.3.2). the set D is dense in H.

It is easy to see that for every polynomial $p \in \mathcal{P}(E)$, the operator $p(A)$ is in the form $p(A) = p(N) + K_0$, where K_0 is a compact operator on H. Let A enjoy the property (α), then we can obtain the following Lemma 3. First of all, since N is an normal operator, it follows that the estimate

$$\|p(A)\|_{\mathrm{e}} = \|p(N) + K_0\|_{\mathrm{e}} = \|p(N)\|_{\mathrm{e}} \leq \|p(N)\| = \|p\|_\infty \tag{1}$$

holds for every polynomial $p \in \mathcal{P}(E)$.

Lemma 3 (Extension of Lemma 2.3.3 in [13]). For fixed vectors $x, y \in H$, a linear functional $\tau_{_{A},x,y} : \mathcal{P}(E) \to F$ is defined by

$$\tau_{_{A},x,y}(p) = \langle p(A)x, y\rangle. \tag{2}$$

Then

(1). $\tau_{_{A},x,y}$ is a bounded linear functional on $\mathcal{P}(E)$;

(2). $\tau_{_{A},x,y}$ is a positive functional on $\mathcal{P}(E)$ if and only if the following inequality

$$\mathrm{Re}\langle p(A)x, y\rangle \leq \|p\|_\infty \langle x, y\rangle$$

holds for every polynomial $p \in \mathcal{P}(E)$.

Proof. (1). By (1), we have

$$\|p(A)\|_{\mathrm{e}} \leq \|p\|_\infty. \tag{3}$$

By definition of the quotient norm, there exists a compact operator K_1 on H such that

$$\|p(A)\| - \|K_1\| \leq \|p(A) + K_1\| \leq \|p(A)\|_{\mathrm{e}} + 1.$$

Thus by (2) and (3)we obtain

$$|\tau_{_{A},x,y}(p)| \leq \|p(A)\|\|x\|\|y\| \leq (\|p\|_\infty + \|K_1\| + 1)\|x\|\|y\|.$$

This shows that $\tau_{_{A},x,y}$ is a bounded linear functional on $\mathcal{P}(E)$.

(2). The proof of (2) is essentially the same as that of lemma 2.3.3 in [13], and is therefore omitted.

Definition 2 ([13], Definition 2.3.4 and Definition 2.3.7). The set of all bounded positive linear functionals τ on $\mathcal{P}(E)$ with norm 1 (i.e., all states) is denoted by $\mathcal{T}\ '$. For each vector $x \in H$, define the set

$$\mathcal{T}_x\ ' = \{y \in H; \tau_{_{A},x,y} \in \mathcal{T}\ '\}.$$

Lemma 4 (Extension of Remark 2.4.4 in [13]). For each vector $x \in D$, the set $\mathcal{T}_x\ '$ is a nonempty closed convex subset of H.

Proof. Using Lemma 3, one can prove the lemma as in [13].

Lemma 5 (Extension of Lemma 2.3.11 in [13]). Every line segment in $\mathcal{T}_x\ '$ has a finite length.

Proof. The proof is completely the same as that of Lemma 2.3.11 in [13].

Lemma 6 (Extension of Lemma 2.4.4 in [13]). If $x \in D$, then for every vector $z \in H$ and every nonnegative integer $k = 0, 1, 2, \ldots$, the following equation

$$\mathrm{Re}\langle A^k((\|P_x z\|^2 - \mathrm{Re}\langle P_x z, z\rangle)x + (I - P_x)z), P_x z\rangle = 0$$

holds, where $P_x z$ is the projection from H to $\mathcal{T}_x{}'$, that is,

$$\|P_x z - z\| = \inf_{y \in \mathcal{T}_x{}'} \|y - z\|.$$

Proof. Write

$$T = A^{*k}, \quad y = P_x z. \tag{4}$$

By definition of the set $\mathcal{T}_x{}'$, the functional $\tau_{A,x,y} = \langle p(A)x, y\rangle$ is a state on $\mathcal{P}(E)$, so that $\tau_{A,x,y}(1) = \langle x, y\rangle = 1$. Set $a = \min\{(\|T\|\|x\|\|y\|)^{-1}, \frac{1}{2}\}$, and define a mapping $\varPhi : (-a, a) \longrightarrow \mathcal{T}_x{}'$ as follows

$$\varPhi(\lambda) = (1 + \lambda \mathrm{Re}\langle Ty, x\rangle)^{-1}(1 + \lambda T)y.$$

Since $0 \le a \le \frac{1}{2}$, the inequality

$$1 + \lambda \mathrm{Re}\langle Ty, x\rangle \ge 1 - |\langle \lambda Ty, x\rangle| \ge 1 - |\lambda||T||x|y| > 0$$

holds for all $\lambda \in (-a, a)$. The same argument as in [13] shows that $\varPhi(\lambda) \in \mathcal{T}_x{}'$ for every $\lambda \in (-a, a)$, so that $\varPhi$ is well defined.

For fixed vector $z \in H$ and fixed nonnegative integer k, the function $\psi : (-a, a) \longrightarrow R^+$ is defied by

$$\psi(\lambda) = \|\varPhi(\lambda) - z\|^2.$$

By properties of the inner product, we have

$$\begin{aligned}\psi(\lambda) &= \left\langle \frac{1}{1 + \lambda \mathrm{Re}\langle Ty, x\rangle}\ (\lambda Ty + y) - z, \frac{1}{1 + \lambda \mathrm{Re}\langle Ty, x\rangle}\ (\lambda Ty + y) - z \right\rangle \\ &= \frac{\langle \lambda Ty + y, \lambda Ty + y\rangle}{(1 + \lambda \mathrm{Re}\langle Ty, x\rangle)^2} - 2\frac{\mathrm{Re}\langle \lambda Ty + y, z\rangle}{1 + \lambda \mathrm{Re}\langle Ty, x\rangle} + \|z\|^2 \\ &= \frac{\lambda^2\|Ty\|^2 + 2\lambda \mathrm{Re}\langle Ty, y\rangle + \|y\|^2}{(1 + \lambda \mathrm{Re}\langle Ty, x\rangle)^2} + 2\frac{\lambda \mathrm{Re}\langle Ty, z\rangle + \mathrm{Re}\langle y, z\rangle}{1 + \lambda \mathrm{Re}\langle Ty, x\rangle} + \|z\|^2.\end{aligned}$$

It is clear that $\psi(\lambda)$ is differentiable on $(-a, a)$ and

$$\begin{aligned}\psi^{'}(\lambda) = 2&\frac{(\lambda\|Ty\|^2 + \mathrm{Re}\langle Ty, y\rangle)(1 + \lambda\langle Ty, x\rangle)^2}{(1 + \lambda \mathrm{Re}\langle Ty, x\rangle)^4} \\ -2&\frac{(\lambda^2\|Ty\|^2 + 2\lambda \mathrm{Re}\langle Ty, y\rangle + \|y\|^2)(1 + \lambda\langle Ty, x\rangle)\langle Ty, x\rangle}{(1 + \lambda \mathrm{Re}\langle Ty, x\rangle)^4} \\ -2&\frac{\mathrm{Re}\langle Ty, z\rangle(1 + \lambda\langle Ty, x\rangle) - (\lambda \mathrm{Re}\langle Ty, z\rangle + \mathrm{Re}\langle y, z\rangle)\langle Ty, z\rangle}{(1 + \lambda \mathrm{Re}\langle Ty, x\rangle)^2}.\end{aligned}$$

Thus we have

$$\psi^{'}(0) = 2\mathrm{Re}\langle Ty, y - z - (\|y\|^2 - \mathrm{Re}\langle y, z\rangle)x\rangle. \tag{5}$$

Noting $y = P_x z$, we see that the the real valued function ψ attains its global minimum at $\lambda = 0$. Consequently $\psi'(0) = 0$. Thus by (4) and (5) we obtain

$$\mathrm{Re}\langle A^k((\|P_x z\|^2 - \mathrm{Re}\langle P_x z, z\rangle)x + (I - P_x)z), P_x z\rangle = 0.$$

Theorem l (Extension of the main result of [13]). If the spectrum of the operator A is real, then A has a nontrivial closed invariant subspace.

Proof. Using Lemma 5 and Lemma 6, one can prove the theorem as in Theorem 2.4.5 of [13].

References

[1] Abramovich, Y.A., Aliprantis, C.D.: An Invitation to Operator Theory. Amer. Math. Soc. (2002)

[2] Aronszajn, N., Smith, K.T.: Invariant subspaces of completely continuous operators. Ann. of Math. 60, 345–350 (1954)

[3] Beruling, A.: On two problem concerning linear transformations in a Hilbert space. Acta Math. 37, 239–255 (1949)

[4] Foias, C., Jung, I.B., Ko, E., Pearcy, C.: Hyperinvariant subspaces for some subnormal operators. Tran. Amer. Math. Soc. 359, 2899–2913 (2007)

[5] Liu, M., Lin, C.: Richness of invariant subspace lattices for a class of operators. Illinois J. Math. 47, 581–591 (2003)

[6] Liu, M.: Invariant subspaces for sequentially subdecomposable operators. Science in China, Series A 46, 433–439 (2003)

[7] Liu, M.: Common invariant subspaces for collections of quasinilpotent positive operators on a Banach space with a Schauder basis. Rocky Mountain J. Math. 37, 1187–1193 (2007)

[8] Liu, M.: On hyperinvariant subspaces of contraction operators on a Banach space whose spectrum contains the unit circle. Acta Math. Sinica 24, 1471–1474 (2008)

[9] Liu, M., Lin, C.: Two operators related to the mohebi-radjabalipour lemma. Acta Anal. Funct. Appl. 10, 97–99 (2008)

[10] Lomonosov, V.I.: On real invariant subspaces of bounded operators with compact imaginary part. Proc. Amer. Math. Soc. 115, 775–777 (1992)

[11] Radjavi, P., Rosenthal, P.: Invariant subspaces. Springer, New York (1973)

[12] Radjavi, H., Troitsky, V.G.: Invariant sublattices. Illinios Journal of Mathematics 52, 437–462 (2008)

[13] Simonič, A.: An extension of Lomonosovs techniques to non-compact operators. Trans. Amer. Math. Soc. 348, 955–975 (1996)

Research of Cooperative Geosteering Drilling Virtual System Based on Network

Xiaorong Gao[1] and Yingzhuo Xu[2]

[1] Institute of Petroleum Engineering, Xi'an Shiyou University, Xi'an, China
gxr_0501@126.com
[2] Institute of Computer, Xi'an Shiyou University, Xi'an, China
yzhxu@xsyu.edu.cn

Abstract. Aiming at the lacks in traditional information revelation and work style for geosteering drilling, a Cooperative Geosteering Drilling Virtual System Based on Network is developed by use of virtual reality and CSCW techniques. Sequentially a cooperative working virtual platform is provided for multi-domain experts and drilling technicians in different locations, which can make them achieve cooperative decision analysis and control while drilling towards drilling operating of the same well at the same time. The three-dimensional visualization of drilling process is implemented in this system using Java 3D. Consequently, it is convenient to control wellbore tracks in real time while drilling, and thus drilling success ratio can be improved. And real-time synchronization between different users' scenes is achieved based on the technique of SOCKET communication and multi-threading. Analyzing the particularity of cooperative operation in geosteering drilling, so a new concurrency control mechanism based on attribute operation in client is introduced to promote the consistency, responsiveness and concurrency of multi-user cooperation. Details are provided about the architecture' design and key techniques of the system in this paper, including the management of drilling virtual objects, constructing of virtual scene, synchronization between multi-users' scenes, concurrency control of multi-user, and so forth.

Keywords: cooperative geosteering drilling; concurrency control; virtual scene; multi-user cooperation; network technology.

1 Introduction

Geosteering drilling, which is the newest high-tech in the drilling industry of the world, could adjust and control well trajectories in real time according to the stratum' s characteristics monitored while drilling to achieve drilling as quickly, safely and accurately. In the process of geosteering drilling, information obtained while drilling is usually inaccurate or fuzzy because of the complex geological condition. In order to use obtained information to achieve control while drilling, firstly these information should be analyzed, processed and revealed in real time, and then it demands that multi-domain experts could collaborate to make a decision and guidance.

In past time, experts usually had to go personally to the well site to work only using the one-dimensional data and all kinds of maps or drawings for interpreting information. However, the details of 3D structure in the reality environment are hard

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 19–28, 2010.

to intuitively show off by one-dimensional data and all kinds of interpreting maps or drawings[1]. In addition, because of the particularity of geosteering drilling, which decision-making and guidance are usually made in filed base and drilling operating is distributed on site far away from the base, it is very hard to organize multi-domain experts to work at well site.

Therefore, we propose a way, which uses Computer Supported Cooperative Work (CSCW) and Virtual Reality techniques to develop a Collaborative Geosteering Drilling Virtual System Based on Network, to visualize geosteering drilling. Meanwhile, this system provides a cooperative working virtual platform for multi-domain experts and drilling technicians in different locations, which can make them achieve cooperative decision analysis and control the drilling operating while drilling towards the same well at the same time through network.

2 System Architecture Design

This system, which is built on a network environment, is the collaborative virtual system for distributed environment. The architecture of system is designed 3-layer (as shown in Figure 1): application layer, control layer and information collection layer.

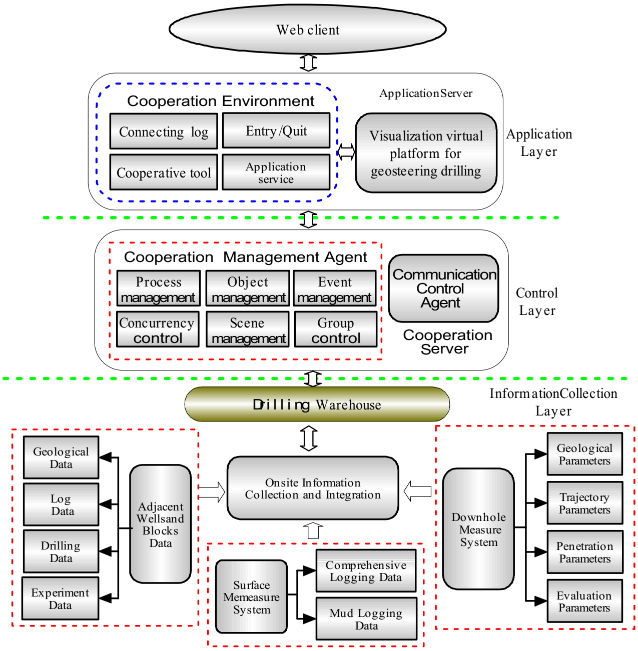


Fig. 1. System Architecture

Application layer: the layer, which consists of a cooperative working environment and a Visualization Virtual Platform for Geosteering Drilling (VVPGD), provides a visualization cooperative drilling virtual environment for users by browser.

1) Connecting log: The module charges of connecting clients and server. It provides kinds of authentication, including initializing authentication, logging authentication and so on.

2) Entry/Quit: The module provides functions for users who want to join an existing workgroup, build a new one, or exit one workgroup. And it could call VVPGD to generate a virtual workspace.

3) Cooperative tool: A subject provides cooperative tools for cooperative drilling, including shared program of application, interactive multimedia, electronic whiteboard, email, etc.

4) Application service: A subject provides services about dealing with kinds of complex problems which include drilling design, drilling monitor, drilling accident diagnosing, as well as the functions of shared information access and query.

5) VVPGD: The platform is a virtual reality system which realizes 3D visualization about drilling objects and dynamic operation process. It also builds a visualization virtual environment of onsite drilling for drilling technicians who can intuitively observe strata, reservoir, well trajectory, target point, and interpret strata and adjust drilling trajectory. Thereby multi-domain experts in different locations could make decision analysis and control onsite drilling through network.

Control layer: the layer manages and controls all the process of cooperative drilling through the cooperation server including Cooperation Management Agent and Communication Control Agent.

1) Cooperation Management Agent: The agent charges of monitoring the system and managing the cooperation process, which consists of 6 modules as following.

a) Process management: The module mainly maintains operation legitimacy during cooperation. As the limitation of network bandwidth, it is hard to get the newest data in all current correlative clients while the data of special object is changed by an operation. For example, one client has deleted an object, whereas another client is modifying it at the same time. Obviously, the later operation is illegal. To ensure the legality of operation, An Entities Owner Table (EOT) there is embedded in this module. This table creates an owner index for every available object. Any client should become the entity’s owner before editing an object. Firstly, the process management module checks the information of entities owner about this object when receiving an operation request from one client. If the operating object is not found in the EOT, the module refuses this. Otherwise, it will match request mark and owner mark, and then decide whether refuse this request or not. When this object hasn’t owners in the EOT, this request user becomes its owner.

b) Object management: The module checks the validity of object, and charges of managing parameters, the changing time of object and the site which objects belong to.

c) Event Management: The module manages users’ events, including detection, analysis, packing and recording. Meanwhile, it filters unimportant information to make sure the transmission of key events, because of the limitation of network bandwidth.

d) Concurrency control: The module realizes the concurrent manipulation of multi-users.

e) Scene management: The module makes sure all users could see the same of scene at the same time. In sever, the information of new object is packed at specified times in order to initialize the state of objects by the package when new users enter into the system. In client, while system initializes local virtual scene, client will download needed resource files and object information to make sure showing the object correctly.

f) Group control: The module realizes the registration and all kinds of managements for cooperative members, which include accounts, grouping, joining in the system and exiting it dynamically. It tracks every member's state while the system is running. At the same time, it provides service about requiring other members' current states for client to realize the cooperated apperceive between members.

2) Communication Control Agent: The agent charges of the management of information routing and Quality of Service, along with bandwidth allocation, storage, transmitting, etc.

Information collection layer: the layer collects and processes the real-time information of Measurement While Drilling, which includes parameters of geology, parameters of wellbore track, parameters of penetration and parameters of drilling evaluation, information of comprehensive logging and mud logging by using of the subsystem of Onsite Real-time Monitoring and Information Processing (ORMIP). And ORMIP synthesizes the results processed and previous adjacent blocks and wells' information, and then transports synthesized information to base through satellite or GPRS for building data warehouse to supply data resources which are used on the upper application layer.

3 Key Techniques of the System Implementation

3.1 Management for Drilling Virtual Objects

Virtual object is the basic unit [2] in the cooperative virtual environment. In this environment, multi-users could operate the same virtual object. In order to realize it, this system abstracts attributes of drilling virtual object, including stratum, well trajectory, target point, and so on. Table 1 illustrates the data structure to adopt for the storage and management of virtual objects.

1) Object Identification: Every virtual object has a unique flag. The unique flag indicates current object while a virtual object is operated or transported.

2) State of Object Location: The real object location is expressed by a 4×4 floating-point matrix, through which the location and rotation angle of the object can be set when system is running.

3) Operation Authority of Object: Some of objects are limited to operate by one user. This group attributes indicate current operators, authority operators' list, whether current object can be operated or not.

4) Activities of Object: This group consists two parts.

a) Current Behavior: It indicates the operation current object is executing.

b) Response Table: It records all responses when an object is operated by a special activity.

5) Other Basic Attributes: There are four parts in this group.

a) Attribute Name (such as color).

b) Current State: It shows the status which the attribute is occupied by users, and this part is used in concurrency control.

c) Current Operation: It indicates current operation type, such as browsing, modification, and so on.

d) History Table: It records history operation on the attribute. It could restore object's states when non-consistency happens.

Table 1. Data Structure of Drilling Virtual Object

Classification	Attribute Name	Data Type
Object Identification	Object ID	Int
	ObjectName	String
State of Object Location	Type	Int
	Transformation	4x4 matrix
	Parent	VirtualObject
	Child	VirtualObject[]
Authority of Object	currentOwner	Int
	floorOwners	Int[]
	controllable	Boolean
Activities of Object	currentBehavior	Behavior
	responseTable	Hashtable
Other Basic Attributes	AttributeName	String
	currentState	String
	currentOperation	String
	historyTable	Event[]

3.2 Building of Virtual Scene for Geosteering Drilling

Virtual scene for geosteering drilling is built by a visualization tool, Java 3D. The graph of scene[3] is a kind of DAG (Directed Acyclic Graph), which defines the geometry, lighting, position, direction, appearance and other visible objects. DAG is created by Java 3D. Implementation process is as follows:

Step 1. Creating an object of Canvas3D used to draw a region that a three-dimensional graphics could be draw in it, this region, which contains the view of in the scene, is rectangular.

Step 2. Creating an object of Virtual Universe to contain the scene to be built.

Step 3. Creating an object of Locale used to put data structures of one group objects. And then, the object of Locale should be related to the object of Virtual Universe. A Local object is a local coordinate system.It provides a citation of one point and determines the orientation signpost of a visualization object in the virtual world.

Step 4. Creating View sub graphs and constructing a view platform for virtual scene in order to observe the whole virtual environment. The view platform needs lots of complicated movement states for watching carefully dynamic scenes. Thus, according to the control data obtained in real time in the process of running, the view platform should fulfill moving. In this system, the platform's moving is realized by using Transform nodes on it, and combining Moving nodes and Alpha to generate view point. The process of creating View sub graphs as following:

Firstly, we need to create a ViewPlatform object, PhysicalBody object and PhysicalEnvironment object. Secondly, a View object should be created. At the same time, it needs to relate with ViewPlatform object, PhysicalBody object, Physicalvironment object and Canvas3D object.

Step 5. Constructing content sub-view and compiling it.

Step 6. Two sub-views above are inserted into Locale nodes.

The basic element of the virtual scene is virtual objects[4]. The process of constructing scene graph for geosteering drilling is as follows: Firstly, all kinds of virtual objects "down hole" generate various nodes in different object coordinate systems, which are parts of whole scene graph. Then, according to space and structure relationship of objects as well as designing parameters[1] in the process of drilling, above nodes are integrated into the world coordinate system. Finally, we could get scenes which we need.

Figure 2 shows the generated 3D view of multi-well tracks and crossed strata. It realizes three-dimensional visualization of the strip of strata adjacent wells by information of drilling and the result of interpreted data, and then well trajectories are embedded in the 3D scene of geology structure. Thereby drilling workers could intuitively watch the trend of well tracks and analyze the kind of crossed stratum. According to this view, it is convenient for controlling well tracks while drilling to achieve drilling exactly into the target oil layer.

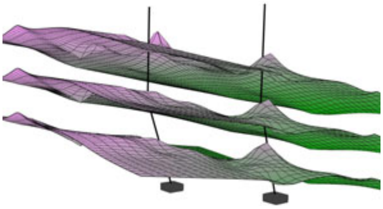


Fig. 2. The 3D visualization view of stratum and well track

3.3 Virtual Scene Synchronization

During the process of collaborative drilling, many cooperative users in different sites could operate the different parts in the same scene by independent operation[5]. In order to allow multi-users to exchange in a virtual scene, we need to ensure that all users have the same scene and synchronization it in time. The synchronization of scenes could be realized by transporting XML data packets using Socket channels.

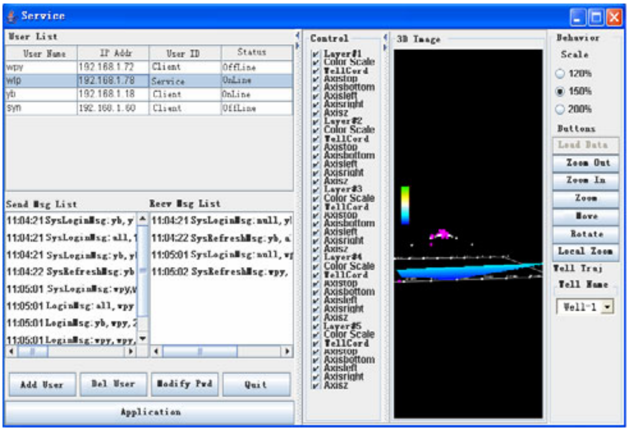
Socket assigns communication endpoint for transporting data between server and client. And server is used of ServerSocket Class[3] to realize it.

```
ServerSocket serverSocket=new ServerSocket(80)•
```

// bingding with port 80

The implementation procedure of scene synchronization is as follows:

Step 1. When the server is running, the Socket interface will be listened. Once there is a request from the client, the Socket connection will be established from the client to server , so the data packet can be transmitted. The server can establish many connections by Multi-Thread at the same time (as shown in Figure 3).



User List

User Name	IP Addr	User ID	Status
wpy	192.168.1.72	Client	OffLine
wlp	192.168.1.78	Service	OnLine
yb	192.168.1.18	Client	OnLine
syn	192.168.1.60	Client	OffLine

Send Msg List

```
11:04:21 SysLoginMsg:yb,y
11:04:21 SysLoginMsg:all,1
11:04:21 SysLoginMsg:yb,y
11:04:22 SysRefreshMsg:yb
11:05:01 SysLoginMsg:wpy,W
11:05:01 LoginMsg:all,wpy
11:05:01 LoginMsg:yb,wpy,Z
11:05:01 LoginMsg:wpy,wpy,
```

Recv Msg List

```
11:04:21 SysLoginMsg:null,yl
11:04:22 SysRefreshMsg:yb,a
11:05:01 SysLoginMsg:null,wp
11:05:02 SysRefreshMsg:wpy,
```

Scale: ○ 120% ● 150% ○ 200%

Well Name: Well-1

Fig. 3. Server interface when multi-users request to connect

Step 2. Set a synchronizing clock on every Socket connection from client to server to make sure it can send the XML data packet of the latest status messages of this client's scene to the server at regular intervals.

Step 3. When the server receives all of the renewed data packets from the online clients, it can acquire the client users' relevant information through interpreting the data packet by JDOM (Java Document Object Model). Then the server will output these users' information, including names, IP address, latest status, and so on, on the monitor-controlling panel (as shown in Figure 3), so the administrator can monitor and manage these users.

Step 4. The server will reconstruct these clients' information and made a new broadcast packet. The broadcast packet will be sent by an independent broadcast thread. To avoid the Socket blocked from sending and receiving, the system sets another thread to send broadcast packet.

Step 5. When the client received the broadcast packet, it will update the position and state of all the virtual objects in the scene through the JDOM's interpretation for the data packet to achieve scene synchronization among the clients.

The server must have Multi-Thread to achieve scene synchronization. So, we have designed four threads on Application Server: ConnectT, ListenT, ExecT and

BroadcastT. Once the ListenT monitored a connecting request from the client, it will start a ConnectT thread to answer, meanwhile start an ExecT thread to run VVPGD and Drilling Services program and connect the Cooperation Server to manage and control it. And it starts BroadcastT thread to broadcast renewed data packet of scene.

1) ConnectT thread: To maintain the Socket connection to the clients; To receive and store XML state renewed data packets from the clients.

2) ListenT thread: To monitor the connection requests from the Socket interfaces; To connect and start VVPGD, Drilling Services program and Cooperation Server; All of these will keep running and forbid the administrator to close the services. Once monitoring a user's request, it will start an ExecT thread and a BroadcastT thread.

3) ExecT thread: To run VVPGD and Drilling Services program; To connect Cooperation Server and transfer parameters.

4) BroadcastT thread: To broadcast the data packet of client's latest updated scene messages to every client at regular intervals.

3.4 Concurrency Control of Multi-user Cooperation

The virtual environment of cooperative drilling allow multi-user operating the shared objects. There will be conflict when several users request to operate one shared object at the same time[6]. We must use a concurrency control mechanism to coordinate, to ensure the consistency, responsiveness and concurrency of the system. So the pivotal thing is to design a rational concurrency control mechanism.

Cooperative geosteering drilling virtual system is different from the traditional multi-user cooperative system[7]. It not only emphasizes the independence of users' operation but also regard the cooperation among the users. The cooperative operation in the system is of particularity as following.

1. User's query operation will be frequent and they want quickly response from the system. Although Query operation is not change any attribute of shared objects it will be use frequently, it will influence the speed of other operations if we don't distinct it with other operations.

For that, we can classify the operations for the shared objects as two types: ① query operation; ② attribute changing operation. Query operation for shared objects will not make conflicts. So, the concurrency control is for the attribute changing operation for shared objects. It will respond at once for the query operation.

2. There will be a conflicts or not when several users operating on the same object at the same time. For example, one user change the color of object O, at the same time the other user change the position of object O, although they operate the same object at the same time, the operations is not conflict. So, the operations for the different attributes of one object will not conflict.

From this we know the system must adopt different concurrency control mechanism. Traditional pessimistic concurrency control mechanism (consistency priority[8]) and optimistic concurrency control mechanism (user's responsiveness priority[8]) have the problem of lower consistency and concurrency to a certain degree[9]. For that, this system proposes a new concurrency control mechanism based on attribute operation in client.

The basic thought is storing state information of shared objects' attributes in client, before users change the shared objects' attribute it will check the attribute's state information: ① If it has not been occupied by other users, the user can operate it and send the request of the attribute's operation token to server; ② If there has a user occupying it then no operation. This avoids view retracement in optimistic concurrency control mechanism to a certain degree, and the responsiveness is better than pessimistic concurrency control mechanism. This mechanism can make users have the maximum live collaboration and make sure the cooperative users have high concurrency and data consistency.

The specific implementation procedure is as follows:

Step 1. When user1 need to operate on attribute A of object O, first of all, determining the state of attribute A of object O in local client, the kinds of state is as follows: ① no one occupies it; ② user1 occupies it; ③ other users occupy it.

Step 2. If the state of attribute A is that no one occupies it or user1 occupies it, user1's operation will be performed immediately and send the operation request to the server.

Step 3. When the server receives the request from user1, it will process as follows according to the current state of attribute A: ① If there is no one occupying it, assign the operation token on attribute A to user1; ② If it is occupied by user1oneself, server will send message of "already get token" to user1;③ If occupied by other users, server will send message of "failing token request" to user1.

Step 4. If user1 receives the failing message, server will update the information of attribute A of object O immediately. Namely, destruction the operation of user1.

When multi users propose operation request to attribute A of object O at the same time, a control strategy of "token + message queue" is adopted in order to avoid conflicting.

Implementation idea: users of different sites send operation request as a message to Cooperative Server, Cooperation Management Agent accumulates them into a message queue by the order of message arriving at Cooperative Server. The Agent also controls a token which is transmitted between cooperative users in a very short time interval. When the token is transmitted to a user, testing whether its operation messages exist in the message queue. If so, the operation token is assigned to the user. The operation is implemented after removing the message from the queue. And the operation result is published to other cooperative members through the network, forming the same operation results. At the same time the token is transmitted to the next user; or directly transmitted to the next user.

Moreover, when user1 possesses attribute A's operation token of object O, to prevent network failure and result in other users can not operate attribute A for long time, server sends inquiry message to user1 in regular time intervals. User1 receives the message and reply to server immediately showing that he is online. If server doesn't received reply after sending inquiry message three times, then the user is defaulted offline, and release of its own token.

4 Conclusion

A new idea and way are put forward to realize cooperative geosteering drilling. The research of this system implements three-dimensional visualization of geosteering drilling process, which could intuitively reveal strata, drilling tracks and so on. And a new cooperative work virtual platform supported by network is constructed, thereby multi-domain experts and drilling technicians in different locations can achieve long distance corporation drilling by the way of intuitive, real-time. Making use of the system can improve the ability of judgment to stratum formation of structure and the control ability of drilling tracks in reservoir, sequentially the controllability of drilling process and drilling efficiency are improved.

Acknowledgment. We are grateful to the Major Scientific and Technological Innovation Project Foundation as well as Natural Sciences Research Foundation of Shaanxi Province for their financial supported under Grant No.2009ZKC02-27 and No.SJ08E$_2$24 for this paper.

References

1. Harding, C., Loftin, B., Anderson, A.: Visualization and Modeling of Geoscientific Data on the Interactive Workbench. In: The Leading Edge, vol. 19, pp. 506–511 (May 2000)
2. Maher, M.L., Gero, J.S.: Agent Models of 3D Virtual Worlds. In: ACADIA 2002, pp. 27–138. California State Polytechnic University, Thresholds (2002)
3. Ming, L., Jingyan, S., Yi, Z.: Virtual Vehicle Simulation System Based on Java 3D Technology. Computer Engineering and Applications 36, 198–202 (2004)
4. Juner, L., Yanning, X., Zhenbo, L., Xiangxu, M.: The Implementation of Multi-user 3D-Scene Using EAI. Journal of System Simulation 14, 1644–1646 (2002)
5. Qingping, L., Low, C.P.: Multiuser Collaborative Work in Virtual Environment Based CASE Tool. Journal of Information and Software Technology 45, 253–267 (2003)
6. Pettifer, S., Marsh, J.: A Collaborative Access Model for Shared Virtual Environments. In: Proceedings of IEEE WETICE 011, pp. 257–272. IEEE Computer Society, Los Alamitos (June 2001)
7. Giertsen, C.A.: Virtual Reality System for Interdisciplinary Petroleum Exploration and Production. Expanded Abstracts with Biographies 1998 Technical Program, Society of Exploration Geophysicists, vol.12,pp. 42-44 (December 1998)
8. Gzara, L.Y., Lombard, M.: Towards a Knowledge Repository for Collaborative Design Process:Focus on Conflict Management. Computers in Industry 55, 335–350 (2004)
9. Xueping, Z., Guofu, Y.: Research on Concurrency Control Mechanism for Collaborative of Virtual Prototype. Application Research of Computers 25, 2959–2961 (2008)

Optimization of Storage Location Assignment for Fixed Rack Systems

Qinghong Wu[1,2], Ying Zhang[2], and Zongmin Ma[1]

[1] College of Information Science and Engineering
Northeastern University
Shenyang, China
[2] School of Electronic and Information Engineering
Liaoning University of Science & Technology
Anshan, China
yingzhang9118@yahoo.com.cn

Abstract. A multi-objective mathematical model and an improved Genetic Algorithm (GA) are formulated for storage location assignment of the fixed rack system. According to the assignment rules, the optimization aim is to maximize the storage/retrieval efficiency and to keep the stability of the rack system. The improved GA with Pareto optimization and Niche Technology are developed. The approach considers Pareto solution sets with the traditional operators, while the Niche Technology distributes the solutions uniformly in Pareto solution sets. The realization of the approach ensures storage location assignment optimization and offers a dynamic decision making scheme for automated storage and retrieval system (AS/RS).

Keywords: Genetic Algorithm; Niche Technology; Pareto optimization.

1 Introduction

It is the first important task to assign the storage location for the goods into and out of the warehouse in the operation of AS/RS. Along with economic gain, the frequency of goods into and out of the warehouse will increase greatly. The plan and shift of warehouse would be convenient, if we use an appropriate warehouse region and a storage location assignment strategy. In order to take and select goods more efficiently, especially to decrease the cost and increase the profit, it is necessary to optimize the goods storage place.

How to optimize the storage place management for the AS/RS rationally is becoming the hot problem paid attention to widely by many people because the efficiency of the AS/RS is mainly decided by the assignment of the warehouse region and goods places. The advantages of the storage location being assigned rationally are as follows. The first is that the work amount of the stowage and the calculating work can be decreased. The second is that the shift distance of goods into and out of the warehouse can be made shorter and the running time can be decreased. The third is that the storage space can be used sufficiently. The fourth is that the stability of the goods rack can be assured and the reliability can be improved [1,2].

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 29–35, 2010.

The best method (Pareto method) is selected from the characteristics of AS/RS multi-objective storage location assignment of the fixed rack system by this paper. In order to obtain approximate solution sets whose distribution property is the best the Niche Technology is adopted. An improved GA with Pareto optimization and Niche Technology are developed. And simulation for the algorithm is researched and the effective of the method used for the location dynamic distribution is approved. Pareto optimization dynamic decision making schemes are offered for decision-makers.

2 An Improved GA Based on Pareto Optimization and Niche Technology

For the part of fixed rack system, when goods quality on the rack and frequency of storage and retrieval are changed, the goods location should be collocated again. The problem is what kind of goods location distribution tactic should be used. And the efficiency of storage and retrieval should be improved to the maximum degree and the stability of the rack should be kept so that the dynamic optimization schemes can be offered for the AS/RS system.

The plane distribution figure of AS/RS fixed rack system is shown as in Fig. 1. There are multi-laneways among the racks. There are two rows of racks on each side of laneway and there is stowage in each laneway. The stowage is busy in and out of the laneway for the running of entry and exit of warehouse. The station that is in and out of the warehouse is set at each gate of the laneway.

In AS/RS the classification storage is used for goods location storage. According to the goods kinds the rack system is divided into some districts. The warehouse district distribution is to collect some goods locations to form goods storage districts. The goods location district can be formed by a rack or some racks, or even a certain goods location in a rack, which can be seen in Fig. 2.

The factors that should be considered in solving the problem of storage optimization are the stability of racks and the goods storage frequency. This paper proposed an improved GA based on Pareto optimization and Niche Technology to solve the problem of dynamic location collocation optimization for the sketch map above.

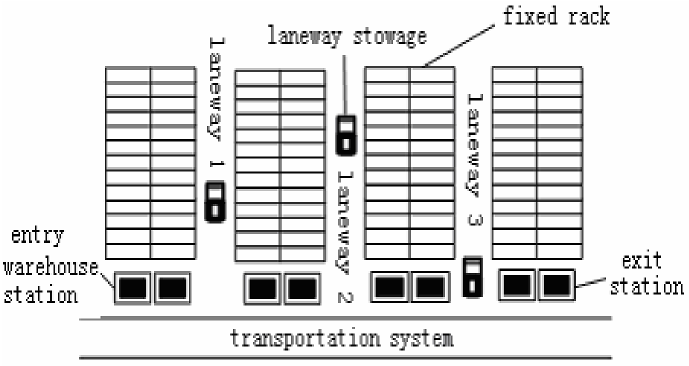


Fig. 1. Sketch map of fixed racks in AS/RS

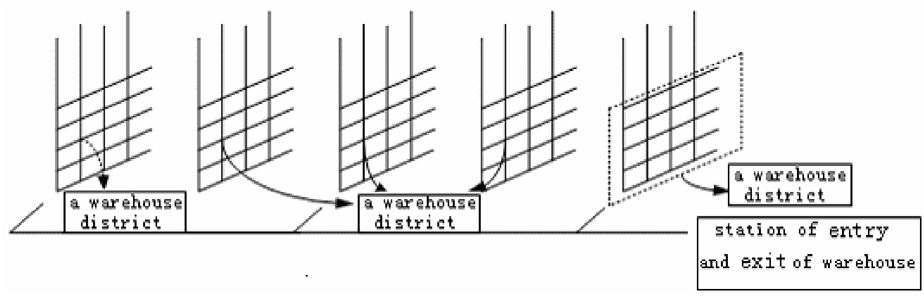


Fig. 2. Sketch map of relationship between warehouse areas and storage locations

3 Niche Technology

In the calculation process of GA, the limited scale of manual colony and stochastic errors brought by arithmetic operators will result in gene excursion, which will bring the same point in the optimization solutions after the colony is combined. The Niche Technology can prevent the gene excursion and make the colony distribute equably in the Pareto optimization solution sets. This paper adopted a better Niche Technology. When the adaptive function of filial generation is super than that of father generation, the father generation is replaced by the filial generation and the replacement is produced when one of the ranks of a filial generation is higher than the highest rank of the father generation, otherwise the father generation will go into the next generation. Under such circumstance, the new filial generation is always better than the father generation. The detail realizing method is that two agents are selected from the father generation and are made across and aberrance. Two new agents go into the father generation and the adaptive function is obtained. Then the adaptive functions of the two new agents are compared with that of the old agents in the father generation.

4 Pareto Solution Sets Filter

The selection procedure cannot assure the optimization properties of each father generation are passed on to the next generation. Because the limited colony scale, some bad points didn't appear again. In the procedure of evolution, some good points only appear one time or twice and then disappear forever. In order to prevent this kind of waste, the definition of Pareto solution sets filter is introduced. Its role is to remain the bad points of each generation and dispose of the bad point in the solution sets.

Suppose n expresses the colony scale, Pop_t expresses the t generation colony, N_r is the number of the ranks, φ is the Pareto solution sets, r expresses the Pareto solution sets scale.

Pop_t is divided into N_r ranks after compositor m subsets are produced, P_1 is the point in *rank*=1, *Pm* is the point in *rank*=*m*, then the Pareto solution sets φ (non-inferior-set) can be expressed as

$$\varphi = \{P_1 | p_{11}, p_{12}, p_{13}, \cdots, p_{1i}, i = 1,2,\cdots r\}$$

In input part, suppose M_{pop} is the colony scale, M_{pareto} is the Pareto solution sets scale, N_r is the colony ranks, P_{si} is the *i*th colony scale, *Gen_max* is the maximum allowed evolution generations. In output part, S_{pareto} is Pareto the solution sets, and the goal function sets is *I*.

The steps of the improved GA based on Pareto optimization and Niche Technology are as follows.

Step 1: the initial colony and external sets are produced when *t*=0.

Step 2: according to the following equations from (1) to (3), calculate the goal function values of each points and the adaptive functions values according to the following equation from (4) to (6) [3].

$$\min \sum_{i=1}^{n} \left| f_i(x) - D_i \right| \tag{1}$$

$$\text{s.t. } x \in S \tag{2}$$

in which D_i is the expectation value of the *i*th goal setup by the decision-maker.

$$S_i = \sum_{j=1}^{M} S(d_{ij})(i = 1,2,\cdots,M) \tag{3}$$

S_i is the sharing function sum of an agent with other agents in the same colony, *M* is the colony scale.

$$\begin{cases} f_{G_x} = G_{x\max} - G_x, G_{x\max} - G_x \geq 0 \\ f_{G_x} = 0, others \end{cases} \tag{4}$$

$$\begin{cases} f_{G_y} = G_{y\max} - G_y, G_{y\max} - G_y \geq 0 \\ f_{G_y} = 0, others \end{cases} \tag{5}$$

$$\begin{cases} f_E = E_{\max} - E, E_{\max} - E \geq 0 \\ f_E = 0, others \end{cases} \tag{6}$$

$G_{x\max}$, $G_{y\max}$ and $E_{\max}$ are the maximums of goal function in late some generations [4].

Step3: according to the definition of non-inferior-point, rank the colony and calculate based on the equation (7).

$$F_i = M_{pop}(N_r - i + 1) / \sum_{i=1}^{N_r} (N_r - i + 1) P_{si} \tag{7}$$

in which, M_{pop} is the colony scale, N_r is the ranks of the colony, P_{si} is the *i*th rank colony scale, F_i is the selection probability of the point in the *i*th rank.

Step 4: make selection, crossover, aberrance, niche and produce new colony.

Step 5: extract the point *Rank* =1 and put it into the Pareto solution sets filter.

Step 6: check the non-inferior-point and eliminate the inferior points. If the number of the points is over the Pareto solution sets scale, delete the redundant points.

Step 7: check whether the solution is convergent, if not, go to the step 2.

Step 8: output the Pareto solution sets and the corresponding goal function values set.

5 Analyze of Experiments

The algorithm above is used to optimize an AS/RS location assignment. There are 13 lines of solid fixed racks, 10 layers and 72 rows, that is, 720 locations in all are assigned in each line rack. In the experiment, L =1m, H =1m, the speed of the stowage V_x =1m/s, V_y =2m/s, the colony scale M_{pop} =80, the maximum allowed evolution Gen_max=220, N_r =10, P_{si} =10 , the crossover probability P_c=0.7, aberrance probability P_m=0.3, the scale of Pareto solution set M_{pareto} =80.

There are 6 kinds of storage goods (W_1-W_6). The improved GA based on Pareto and Niche Technology is used to optimize assignment to some row of the rack. According to the goods kinds and average work frequency, the goods location after optimization is divided into 5 areas I-V shown in Table 1, in which II is the kind of W_5 whose frequency is the maximum (0.25), the percent of the assignment location is 20%. Goods W_6 is assigned in IV area, goods W_4 in V area, and goods W_2 and W_3 assigned in I area, goods W_1 in III area.

In Table 3, location assignment for goods W_5 whose codes is from 1 to 18 is given and the result showed that the location can be adjusted by used the algorithm in this paper according to the requirement of goods in each period and the goods weight. The algorithm satisfied the requirement of assuring the goods stability and high efficiency of restoring and retrieval.

Table 1. Storage rack areas table by product class

Goods kinds	Average work frequency	Assignment goods areas	Percent of location assignment (%)
W_1	0.07	III	9
W_2	0.11	I	23
W_3	0.11	I	4
W_4	0.13	V	14
W_5	0.25	II	19
W_6	0.18	IV	21

Table 2 is the compare of methods Random place assignment tactic, Special subarea assignment tactic and Assignment tactic in this paper for the location assignment of some row of fixed rack. The result showed that the tactic in this paper is more idea than other methods.

Table 2. Performance compare of different allocation strategies

Strategies	Gravity center of rack (m)		Work time (s)
	Gx	*Gy*	
Random place assignment tactic	26.12	3.02	1846
Special subarea assignment tactic	21.00	1.65	1105
Assignment tactic in this paper	17.62	0.89	536

Table 3. Area products of W5 class storage location allocation table

Goods code	Average requirement/ work period	Goods weight (kg)	Random Storage place (i, j)	Storage place afteroptimization (i, j)
1	4	9	(1,2)	(13,6)
2	5	21	(3,5)	(5,1)
3	8	3	(3,7)	(3,6)
4	14	7	(12,7)	(15,7)
5	6	11	(5,3)	(15,8)
6	15	6	(5,10)	(2,2)
7	2	8	(2,3)	(3,3)
8	15	12	(11,6)	(4,6)
9	4	4	(9,7)	(14,7)
10	2	6	(12,4)	(7,1)
11	2	5	(8,6)	(14,5)
12	8	8	(6,8)	(8,3)
13	5	23	(2,5)	(2,1)
14	2	4	(12,3)	(7,2)
15	13	7	(4,8)	(16,2)
16	2	3	(11,4)	(7,3)
17	2	4	(1,5)	(18,2)
18	4	15	(12,8)	(8,1)

6 Conclusions

This paper designed an improved GA based on Pareto optimization and Niche Technology to optimize the location of fixed rack and approved the validity of the algorithm by experiments. After using the method in the paper, the gravity center is lower and the time of restoring and retrieval goods is shorter than those by using other methods. The results showed that the algorithm proposed in this paper is effective and can be used widely in practice.

References

1. Hsieh, S., Tsai, K.C.: A BOM oriented class-based storage assignment in an automated storage / retrieval systems. Int. J. Adv. Manuf. Technol. 17, 683–691 (2001)
2. Thonemann, U.W., Brandeau, M.L.: Optimal storage assignment policies for automated storage and retrieval systems with stochastic demands. Management Science 44(1), 142–148 (1998)
3. Ulungu, E.L., Teghem, J., Fortemps, P.H., et al.: MOSA method: a tool for solving multi-objective combinatorial optimization problems. Journal of Multi-criteria Decision Analysis 8, 221–236 (1999)
4. Thierens, P., Bosman, D.: The balance between proximity and diversity in multi-objective Evolutionary algorithms. IEEE Transactions on Evolutionary Computation 7(2), 174–188 (2003)

Evaluation Query Answer over Inconsistent Database with Annotations

Aihua Wu

Dept. C.S. of Shanghai Maritime University, Shanghai 201306, China
061021058@fudan.edu.cn

Abstract. In this paper, we introduce an annotation based data model of relational database that may violate a set of functional dependency. In the data model, every piece of data in a relation can have zero or more annotations with it and annotations are propagated along with queries from the source to the output. With annotations, data in both input data and query answer can be divided into certain and uncertain part down to cell level. It can avoid information loss. To query an annotated database, we propose an extension of SPJ-UNION SQL, *CASQL*, and algorithms for evaluating *CASQL* so that annotations can be correctly propagated as the valid set of functional dependency changes during query processing. Last, we present a set of performance experiments which show that time performance of our approach is acceptable, but performance in information preserving is excellent.

1 Motivation and Introduction

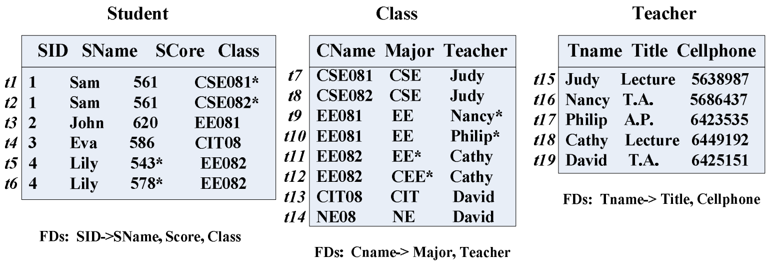
Data is consistent if it satisfies all integrity constraints predefined on it. Inconsistency, a type of invalidity, is viewed as a serious deficiency of any type of data, and should be avoided. Unfortunately, there are situations, such as data merging, data exchange, and etc, where integrity constraints may be violated unavoidably.

When database is inconsistent, how do we give a "clean" answer to user query? Most previous researches focus on strategies of data cleaning[1-4], and consistent query answer(CQA)[5-7]. Data cleaning try to correct errors in the data so that "clean" answer can be computed against the "clean" database. It is helpful in some applications, but it requires user's interference, and can not guarantee correctness of the database when insertion or modification is used. CQA technique defines consistent query answer as the common part of answers to the query on all repairs of the inconsistent database. It is useful to find rules for computing sure query answers, but it may also lead to information loss like data cleaning with deletion. [1]

In [8], we present an annotation based approach for answering queries over relational databases that may violate a set of functional dependencies (FDs). In this approach, every piece of inconsistent data is attached with one or more annotations. The approach separates certain data from uncertain data down to attribute level, and loss nothing. With annotations, user can know whether a piece of data is credible, and even judge which of the conflicting values is correct with his background knowledge.

[1] Supported by the program of Shanghai Municipal Education Commission under Grant No.09YZ243.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 36–43, 2010.



Student

	SID	SName	SCore	Class
t1	1	Sam	561	CSE081*
t2	1	Sam	561	CSE082*
t3	2	John	620	EE081
t4	3	Eva	586	CIT08
t5	4	Lily	543*	EE082
t6	4	Lily	578*	EE082

FDs: SID->SName, Score, Class

Class

	CName	Major	Teacher
t7	CSE081	CSE	Judy
t8	CSE082	CSE	Judy
t9	EE081	EE	Nancy*
t10	EE081	EE	Philip*
t11	EE082	EE*	Cathy
t12	EE082	CEE*	Cathy
t13	CIT08	CIT	David
t14	NE08	NE	David

FDs: Cname-> Major, Teacher

Teacher

	Tname	Title	Cellphone
t15	Judy	Lecture	5638987
t16	Nancy	T.A.	5686437
t17	Philip	A.P.	6423535
t18	Cathy	Lecture	6449192
t19	David	T.A.	6425151

FDs: Tname-> Title, Cellphone

Fig. 1. An inconsistent database with annotations(cell value with "*" is inconsistent)

Based on a database with inconsistency annotations, we discuss in this paper the method for user to submit proper queries and get query answers whose inconsistent piece of data are correctly annotated. We proposed an extension of SQL, *CASQL*, with which user can specify FDs of input relation schema, domain equality, and whether the input relation instance is annotated. We also describe how we compute annotations for a given query, so that annotations can be correctly propagated, and query answer is valid and sound w.r.t. both user query and semantic of data consistency in algorithms for evaluating *CASQL*. Experimental evaluation shows that the approach may bring acceptable extra cost, while using the annotation system can avoid information losing in answering queries over inconsistent database.

Our Contribution This paper makes the following technical contributions:

- We propose an annotation system for answering queries over inconsistent database. We define two types of annotations and three types of operations on them. We also study the relational operations over annotated relations and give them new semantics. We present algorithms for computing valid annotations and FD set in query answers for a SELECT-PROJECTION-JOIN-UNION query with arithmetic comparisons. Furthermore, we propose an extension of SQL, *CASQL*, which provides a way to query annotated relations.
- We present a performance study of annotation computing, constraints computing and query evaluation. We consider different degrees of inconsistency and database sizes to test the applicability of our approach. The experimental evaluation shows that time performance of our framework is acceptable.

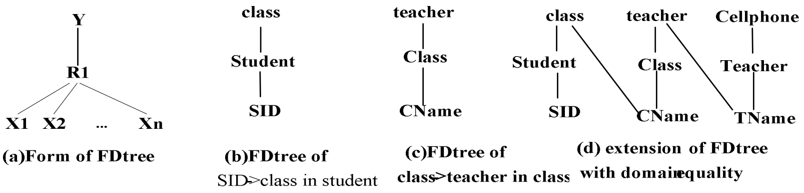
Organization This paper is organized as follows. Section 2 defines some basic concept, including our annotations and its operations. Section 3 presents CASQL and its evaluation algorithm. Section 4 is experimental results. The last is conclusion.

2 Basic Conception

Definition 1 ***domain equality (DEQ)*** Given database schema *D*, domain equality statement *X=Y* is true iff for every tuple *t* of *D*, *t[X]=t[Y]*.

Definition 2 ***FDtree*** Given a database schema *D*, *suppose X->Y* is a FD on *R* of *D*, FD tree of *X->Y* over *D* is an unordered tree in form of fig.2(a) whose nodes can only be attribute name or relation name of *D*: 1) *Y* is the root of the tree, 2) each member of *X* is a leaf of the tree, 3) relation *R* is a node connecting X and Y, and 4) domain equal attribute can directly connect to attribute with arc -. And we name attributes in the left side of a FD *independent* attributes and the right side *dependent* attributes.

Example 1 FD tree of part FDs over database of fig.1. is shown in fig.2.



(a)Form of FDtree

(b)FDtree of SID->class in student

(c)FDtree of class->teacher in class

(d) extension of FDtree with domain equality

Fig. 2. FD trees

In FDtree, any two domain equal attribute nodes can be viewed as a substitution of the other. Algorithm for calculating implied FDs are shown in section 3.2.

Definition3 annotated relation: Given a relation *R* and its FD sets *F*, if all inconsistent piece of data of *R* is marked with annotations, *R* is a relation annotated w.r.t. *F*.

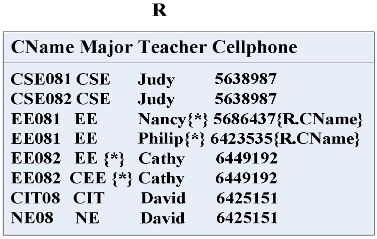
In annotated relation, certain piece of data has no annotation while uncertain piece of data can have one or more annotations with it. For each piece of data, we use a list to record annotations with it. There are two situations that inconsistent piece of data can be implied: data that is inconsistent in initial input database and data that is only inconsistent in the query result. We use mark * for the former, and *independent* attribute name for the latter. Correspondingly, we call mark * *static inconsistency mark* and the latter a *dynamic inconsistency mark*

In annotated relation, tuple can be viewed as combination of value and mark. In rest of this paper, *t* is tuple with mark, *tv* denote to tuples without regard to marks on it, and $t[X]^{mark}$ represent mark list on attribute *X* of *t*. For any tuple *A* and *B* of an annotated relation, 1) if a) $Av=Bv$ and b) for any attribute *X*, $A[X]^{mark} = B[X]^{mark}$, we say *A* equal to *B*, denoted as $A=B$; else if $Av=Bv$ but $A \neq B$, we say that *A* is value equal to *B*, denoted as $A\,\underset{=}{v}\,B$; 2) if a) $Av=Bv$ and b) for any attribute *X*, if *A.X* is marked, *B.X* must be marked, *A* is mark-included in B.

During the query evaluation, new FDs may cause new dynamic inconsistency mark, while deleting FDs can cause removing of a dynamic inconsistency mark. Mark deducing and mark dividing, two type of operations on annotations, are presented here for the above two type of transformations.

Mark deducing Let *D* be an annotated database, *e* be a query expression over D. $\forall(X->Y)((X->Y)\in Drv(e(D))^{\wedge}(X->Y)\notin Drv(D))$, For any tuple *t1,t2* ($t1 \neq t2$) in *e(D)*, if *t1[X]=t2[X]* while $t1[Y] \neq t2[Y]$, "e(D).X" will be added into $t1[Y]^{mark}$ and $t2[Y]^{mark}$. This operation is called *mark deducing*. Rules for Drv(D) and Drv(e(D)) are presented in [9].

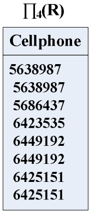
Mark dividing Let *D* be a database whose relations are annotated, let *e* be a query expression of projection over *D*. For any tuple *t* and any attribute *X* of *e(D)*, for any



R

CName	Major	Teacher	Cellphone
CSE081	CSE	Judy	5638987
CSE082	CSE	Judy	5638987
EE081	EE	Nancy{*}	5686437{R.CName}
EE081	EE	Philip{*}	6423535{R.CName}
EE082	EE {*}	Cathy	6449192
EE082	CEE {*}	Cathy	6449192
CIT08	CIT	David	6425151
NE08	NE	David	6425151

Fig. 3. Mark deducing in evaluation of Q1
Q1: select Class.*, Cellphone
from Class, Teacher
where Class.Teacher = Teacher.TName



Π_4(R)

Cellphone
5638987
5638987
5686437
6423535
6449192
6449192
6425151
6425151

Fig. 4. Mark dividing in evaluation of Q2
Q2: Select cellphone from R

dynamic mark Y of *t[X]*, if attributes of *e[D]* can belong to one relation, and Y isn't an attribute of *e(D)*, Y will be deleted from $t1[X]^{mark}$.

3 CASQL and Its Evaluation

In this section, we focus on the problem of how to query annotated database. At present, we only consider SPJU queries.

CASQL(Certain Answer with SQL) is an extension of a fragment of SQL that corresponds to conjunctive queries with union and arithmetic comparisons. *CASQL* extends the Select-Project-Join-Union fragment of SQL in two ways: 1) a FD clause to allow users to specify sets of input functional dependency and domain equality, and 2) a BASIC predication in FROM clause to distinguish an annotated relation and an unannotated relation so that different action will be taken during the query evaluation.

Definition 4 A CA*SQL query* is a query of the form Q1 UNION Q2 UNION,… , UNION Qn where each Qi(1<=i<=n) is a *query fragment* of the form as below:

SELECT [DISTINCT] selectlist
FROM *fromlist*
[WHERE *wherelist*]
[FD *FDset1* OF *r1* [, *FDset2* of *r2*, …] [WITH *DEQExpr1* [, *DEQExpr2*, …]]]

Here clauses or predications between "[" and "]" is optional.

- The *fromlist* is in form of " [BASIC] R1 r1, [BASIC] R2 r2,…" where the predication BASIC before each relation indicates the relation is an unannotated one. We don't consider views in *fromlist*. Furthermore, if keyword BASIC appears before a relation *r*, FDs of r must be present in FD clause.
- The *selectlist* is in form of "r1.A1 [AS D1], r2.A2 [AS D2],…" where *Ai* is an attribute of *ri*. Functions and arithmetical expressions are not considered.
- The *wherelist* is a conjunction of one or more arithmetic comparisons between attributes of relations or between attributes of relations and constants. And all the conditions don't restrict annotations or domain, but value of tuples. Domain equality will be specified by FD clause.
- The FD clause specifies set of FD of each relation in *fromlist* and DEQ between them. FDs are specified by clauses of "FDseti OF ri " where *FDseti* is a set of

FD and ri is also a relation in *fromlist*. DEQs are specified in form of "ri.Am=rj.An" where Am, An are attributes of relation ri and rj respectively. We don't consider domain equlity between two attributes from one relation.

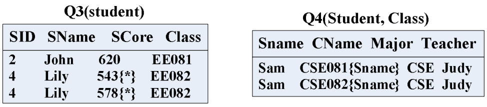
Three levels of operations is implied in a CASQL query. The first level is query operation on tuples which is similar to traditional SQL query. The others are computation and propagation of valid set of FDs and annotations with the query evaluation. And we will present and illustrate query processing and semantic details for a selection, projection, join and union CASQL query in the next section.

We present details of how we evaluate different *CASQL* in the following.

```
-----------------------------------------------
Algorithm  Query evaluation of CASQL
Input: CASQL SYNTAX TREE casql
Output: query result of casql
1.  set FDset=all FDtrees in casql; set DEQs=all DEQs in casql.
2.  For each r in Fromlist
    If "basic" or "BASIC" appear before r
    {set fdr=FDs of r
     basicAnnotating(r,fdr);//annotated r with * w.r.t. fdr}
3.  Represent casql as query tree (QTree) like what we do with SQL query and use optimizing
    rules for SQL query to get query optimized QTree.
4. ExcuteQuery(QTree, FDset, DEQs)
----------------------------------------------
Function ExcuteQuery( query tree Qtree, set of valid FD in all leaf node relations FDset, set of
DEQ between all leaf node relations DEQs)
{  CurrentFD=NULL, inputDEQs=NULL
   scan QTree in postOrder sequence, and for each operation node N of QTree (in the rest,
   relation corresponds to node N are denoted N)
   {  if N->leftChild is leafNode
         currentFD=currentFD $\cup$ (fdr1=all FDs of N->leftChild)
      if N->rightChild is leafNode
         currentFD=currentFD $\cup$ (fdr2=all FDs of N->rightChild)
      inputDEQs= all DEQs between relations of CurrentFD
      Switch (N.operationType)
      {case SELECTION:
        a)  Select tuples who satisfy the condition and tuples who are inconsistent on condition
            attribute from N.child into N.
        b)  Generate name for relation of the query result with a global incrementor, and record
            provenance of each attribute.
        c)  If inputDEQs!=NULL
           {newFD=CalculateNewFDs(currentFD, inputDEQs);
            if newFD!=NULL
            {N'=select tuples from N.child-N who are equal to a tuple in N on keyatt.
             do mark deducing in N but with data from N+N' w.r.t. all FD in newFD}
              currentFD=currentFD $\cup$ newFD}
      Case PROJECTION:
        a)  Project tuples from N.child into N.
        b)  Generate name for relation of the query result with a global incrementor, and record
            provenance of each attribute.
        c)  currentFD=calculate valid FDs use rules in section2 with currentFD
        d)  do mark dividing in N and delete tuples equal to or mark-included in another tuple
            of the query result.
```

```
  Case JOIN:
    a) Join those t from N->leftChild and s from N->rightChild into N that t and s satisfy
       join condition in value or s is inconsistent on join attributes.
    b) Generate name for relation of the query result with a global incrementor, and record
       provenance of each attribute.
    c) If inputDEQs!=NULL
        { newFD=CalculateNewFDs(currentFD, inputDEQs);
          if newFD!=NULL
          {do mark deducing according to all FD in newFD in N }
          currentFD=currentFD ∪ newFD}
  Case UNION:
    a) Union all tuples from N->leftChild and N->rightChild into N, but erase tuples that
       are equal to or mark-included in another tuple.
    b) Generate name for relation of the query result with a global incrementor, and record
       provenance of each attribute.
    c) Do consistency checking in N and mark all unannotated uncertain piece of data
       w.r.t currentFD. }}}
---------------------------------------------------
  Function CalculateNewFDs(FDtrees FDs, DEQs DE)
  {  Set s=s'=FDs  //for set of all functional depencencies
     add equal FDs to s with substitution
     s join with s to find new dependency
     remove duplicate FDs from s
     Output s-s' }
--------------------------------------------------------
```

Q3: select * From basic Student Where SCore>=600
    FD {Sid->SName,Sid->SCore, Sid->Class} of Student
Q4: select SName,CName,Major,Teacher from Student, basic Class
    where SName='Sam' and Class=CName
    FD {Sid->SName,Sid->SCore, Sid->Class} of Student,
        {CName->Major, CName->Teacher} of Class with Class=CName



Q3(student)

SID	SName	SCore	Class
2	John	620	EE081
4	Lily	543{*}	EE082
4	Lily	578{*}	EE082

Q4(Student, Class)

Sname	CName	Major	Teacher
Sam	CSE081{Sname}	CSE	Judy
Sam	CSE082{Sname}	CSE	Judy

Fig. 5. Certain query answer for Q3, Q4 over database in fig.1

4 Experiment Evaluation

Experimental environment Intel Celeron 420 1.6GHZ CPU, 512MB memory ; windows XP+SP2. VC++ 6.0, and SQL Server 2000.

Data set generation. To test the efficiency of our techniques on very large data sets, we used a synthetic data generator which can be run with two parameters, database size and dirty ratio.

We produce two groups of database in schema of database shown in fig.1. The first group of databases is in size of 0.1GB, 0.5GB, 1GB and 2GB whose dirty ratio is 1%, and the second group of databases is in size of 0.5 GB but with different dirty ratio of 5%, 10%, 15% and 20%. All the data sets are sorted by key attribute. And proportions of data size of table *Student, Class* and *Teacher* is 60%, 25% and 15% respectively.

Queries. We use 7 *CASQL* queries in the experiment. Queries q1-q4 are about one table, q5 and q6 are join query, q7 is a union query. All input table are *basicAnnotated*. All the queries are shown below in form of algebra expressions.

q1: $\sigma(\mathrm{Student})$ **q2:** $\sigma_{Class='\ CSE053'}(\mathrm{Student})$ **q3:** $\prod_{2,3,4}(\mathrm{Student})$ **q4:** $\prod_{2,3,4}(\sigma_{Class='\ CSE053'}(\mathrm{Student}))$

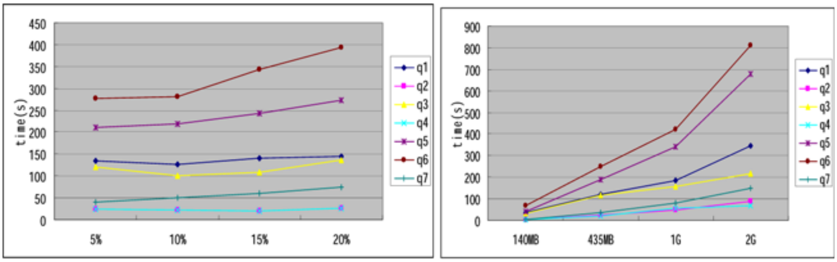
q5: $\prod_{2,3,6}(\mathrm{Student}\infty\mathrm{Class})$ **q6:** $\prod_{2,3,6,7,9}(\sigma_{SCore>=600}(\mathrm{Student})\infty\mathrm{Class}\infty\sigma_{title='\ reofessor'}(\mathrm{Teacher}))$

q7: $\sigma_{Class='\ CSE053'}(\mathrm{Student})\ \mathrm{U}\ \sigma_{SCore>=600}(\mathrm{Student})$

Storage of annotations. We create annotation table for each table. Each tuple and its annotation tuple share same identifier. Both tables are sorted in same sequence.

CQSQL. We first executed all the queries on different scales of annotated database with 1% dirty ratio. The results are shown in fig.6(b). As expected, the execution time of each query increases in proportion to the increase of dataset scale. That is because most of the time is spent in reading and writing data and annotations from the database, while computation of annotation is relatively unimportant in the total time.

Furthermore, the execution time goes more quickly as more tables are joined together. In fact, new FDs goes sharply as input FD set and DEQ set goes up when more table joined together, and *mark deduing* is the largest time consuming annotation operation.



(a) performance of 582MB database (b) performance when dirty ratio is 1%

Fig. 6. Performance comparisons of different queries

Table 1. comparison of information lost for our system with repairing by deletion and CQA

	q1			q4			q8($\prod_{sname,class}$(q4)		
	RWD	CQA	CASQL	RWD	CQA	CASQL	RWD	CQA	CASQL
140MB-0.01DR	0.75%	0.75%	0	33.3%	33.3%	0	33.3%	0	0
435MB-0.01DR	0.75%	0.75%	0	33.3%	33.3%	0	33.3%	0	0
1GB-0.01DR	0.75%	0.75%	0	33.3%	33.3%	0	33.3%	0	0
2GB-0.01DR	0.75%	0.75%	0	33.3%	33.3%	0	33.3%	0	0
582MB-0.05DR	3.75%	3.75%	0	33.3%	33.3%	0	33.3%	0	0
582MB-0.10DR	7.50%	7.50%	0	33.3%	33.3%	0	33.3%	0	0
582MB-0.15DR	11.25%	11.25%	0	33.3%	33.3%	0	33.3%	0	0
582MB-0.20DR	15.00%	15.00%	0	33.3%	33.3%	0	33.3%	0	0

Secondly, we executed all the queries on 582MB annotated database with different dirty ratio. The results are shown in fig. 6(a). The execution time of q1- q4 and q7 differ little. It's because that the five queries are applied over one table and no need to update annotations. Performance of q5 and q6 show that more tables are joined together, more execution time is needed.

Information Loss. We analyze the performance of information loss of CQA and database repairing with tuple deletion (RWD in next), and compare them with our experiment result of query q1, q4 and a projection on q4 (q8, detail shown in table1) over the two group datasets. Information loss rate are calculated as total number of all lost cells divide total number of cells satisfying the query. The experimental result shows that our method loss nothing.

5 Future Work

We introduce an annotation based data model of relational database that may violate a set of FD. The data model doesn't filter any query-satisfied data from the answer, but distinguish them with annotations. With annotations, data in both input database and output query answer can be devided into certain part and uncertain part down to attribute level. Based on the data model, we propose CASQL language, so that for any different SPJ-Union query, the approach can correctly compute and propagate annotations to the query answer. It can avoid information loss. Insofar, our approach are limited to constraint type of FD and SPJ-Union queries with arithmetic comparisons. As a future work, we will extend the method to aggregation queries and complex constraints.

References

1. Bohannon, P., Flaster, M., Fan, W., Rastogi, R.: A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification. In: SIGMOD Conference 2005, pp. 143–154 (2005)
2. Wijsen, J.: Database Repairing using Updates. ACM Transactions on Database Systems 30(3), 722–768 (2005)
3. Lopatenko, A., Bravo, L.: Efficient Approximation Algorithms for Repairing Inconsistent Databases. In: ICDE 2007, pp. 216–225 (2007)
4. Franconi, E., Palma, A.L., Leone, N., Perri, S., Scarcello, F.: Census data repair: a challenging application of disjunctive logic programming. In: Nieuwenhuis, R., Voronkov, A. (eds.) LPAR 2001. LNCS (LNAI), vol. 2250, pp. 561–578. Springer, Heidelberg (2001)
5. Arenas, M., Bertossi, L.E., Chomicki, J.: Consistent query answers in inconsistent databases. In: Proceeding of the PODS Conference,Philadelphia, pp. 68–79 (1999)
6. Chomicki, J.: Consistent Query Answering: Five Easy Pieces. In: Schwentick, T., Suciu, D. (eds.) ICDT 2007. LNCS, vol. 4353, pp. 1–17. Springer, Heidelberg (2006)
7. Bertossi, L., Chomicki, J.: Query Answering in Inconsistent Databases. Logics for Emerging Applications of Databases, 43–83 (2003)
8. Wu, A., Tan, Z., Wang, W.: Annotation based query answer over inconsistent database. Journal of Computer Science and Technology (JCST) 25(3), 467–479 (2010)
9. Klug, A.C.: Calculating Constraints on Relational Expressions. ACM Trans. Database Syst. 5(3), 260–290 (1980)

Research and Implementation of Index Weight Calculation Model for Power Grid Investment Returns

Wei Li[1], Guofeng Chen[1], and Cheng Duan[2]

[1] School of Control and Computer Engineering, North China Electric Power University
liwei@ncepu.edu.cn,
chengf353@163.com
[2] College of Electrical and Electronic Engineering, North China Electric Power University
Beijing, China
duancheng1985@hotmail.com

Abstract. Investment evaluation of electric power grid is an evaluation of proportional relation between the profit, which is gained from investment money of electric power grid enterprise in a certain period, and the investment which the profit occupy or consume. The investment evaluation of electric power grid is reflected by the index evaluation system, the importance of which is reflected by index weights, the precise calculation of the weights have a very important role. In this article, we have established a weight calculation model, the coefficient of variation method, Delphi method, entropy method are used to calculate the weight, and finally combination of three methods to calculate the weight, get the Combination weights.

Keywords: Investment Evaluation, Coefficient of variation method, Delphi method, Entropy Method, Combined weights, EJB Technology, J2EE.

1 Introduction

For a project in any environment , in order to study the effect of its operations, it must be reasonable assessment and evaluation, the most common approach of assessment and evaluation is to establish evaluation index system, calculation of the index weights, then calculated the index overall score and show. Power Grid Project is also assessing in this way. Power grid enterprises as a basic infrastructure projects, the significance of protection social and economic development are enormous. The foundational status and development of power grid enterprises determine them need appropriate investment.

Establish power grid evaluation index system to reflect the power grid investment income is a reasonable way. The index system, which wide coverage, can objectively reflect the relationship between investment and returns, and can reflect the influence of various factors of power grid investment efficiency from different angles.

Evaluation System is an organic whole. We need to specify the weight of each index in the establishment of evaluation index system, which to reflect the importance of the index.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 44–52, 2010.

2 Index Weight Calculation Model

All index of the index system reflect the investment returns of the power grid enterprise from different angles. Weight calculation model can put the method of measure the importance of each index into a comprehensive quantitative results through a mathematical method. The model combined a variety of algorithm to calculation results. Delphi method, Analytic Hierarchy Process, Coefficient of variation method, entropy method and other methods are common methods of weight calculation[1]. Delphi method and AHP are subjective weight calculation method. We take advantage of Analytic Hierarchy Process to obtain the index weight by using mathematical methods melt into expert's advices, which both subjectivity and objectivity, However, in practice, we need for training of experts, or may be cause can not pass the consistency test; Delphi method has intuitive features, and suitable for the actual operation. Coefficient of variation method, Entropy method and other methods are objective weight calculation method, Variation coefficient method based on the variation size of observation value to assign weight for evaluation object, and assumption that a weak correlation between the index. Entropy method is based on the amount of information content of all index transmitted to the decision makers to determine the index weight, which can reflect the information content of index itself[2]. Therefore, we used the Delphi method, Coefficient of variation method, Entropy method in the weight calculation model, finally an integrated combination weights obtained from above three ways through Combination weighting method[3].

3 Index Weight Calculation

3.1 Coefficient of Variation Method

Coefficient of variation is a statistic data which commonly used measure the data difference. Its basic principle is that the greater the variation of the index, the greater the impact on the evaluation. The weight reflects the resolution capability of index. In order to avoid the impact of index different dimension and magnitude, Coefficient of variation method use the treatment value as the weight of each index directly.

(1) Suppose there are m evaluation index of investment returns, and n regions of power grid evaluation, X for the original data matrix, X_{ij} is the j index value of the i objects.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{1}$$

(2) Calculate the standard deviation of each index, to reflect the absolute variation of the index.

$$S_j = \sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2 / n} \tag{2}$$

S_j in the formula is the standard deviation of the index j.

(3) Calculate the coefficient of variation of each index, to reflect the relative variation of each index.

$$U_j = S_j / \bar{x_j} \tag{3}$$

(4) Normalized coefficient of variation of each index, to obtain the weight of each index.

$$w_j = U_j / \sum_{j=1}^{m} U_j \tag{4}$$

3.2 Entropy Method

Entropy method determine the index weight according to the information content that transmit to the decision makers by each index. The greater the difference of an index, the smaller the entropy value, the more information be contained and transmitted by index, the greater the corresponding weight. According to the definition of information theory, the information content of the i signal in information transmission channels $L_i = -\ln h_i$, h_i is the probability of this signal. Therefore, if there are n signals, the probability of their occurrence is $h_1, h_2, h_3 \cdots\cdots h_n$, then the average information content of all n signals, Scilicet Entropy is $-\sum_{i=1}^{n} h_i \ln h_i$.

Entropy method is divided into the following steps:

(1) Assumed there are n power grid regions need to evaluation and m index of investment returns have to set weights. X_{ij} is the preset value of the sample $i(i \leq n)$relative attribute $j(j \leq m)$, and form the primitive data matrix $X = (x_{ij})_{n\times m}$.

(2) Get the optimal value of each index x_j , if the j is positive index, x_j the bigger the better, if the j is inverse index , x_j the smaller the better.

(3) Define the proximity of x_{ij} relative to x,

$$F_{ij} = \begin{cases} x_{ij}/x_j & x_j = \max\{x_{ij}\} \\ x_j/x_{ij} & x_j = \min\{x_{ij}\} \end{cases} \tag{5}$$

and get the matrix $F = (F_{ij})_{n\times m}$

(4) Normalization to F_{ij}:

$$f_{ij} = F_{ij} / \sum_{i=1}^{n}\sum_{j=1}^{m} F_{ij} \tag{6}$$

Suppose $0 \leq f_{ij} \leq 1, \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = 1$,and get the matrix $f = (f_{ij})_{m\times n}$.

(5) Calculate the conditional entropy of index j,

$$E_j = -\sum_{j=1}^{n}({}^{f_{ij}}/_{f_j})\ln({}^{f_{ij}}/_{f_j}) \tag{7}$$

suppose $f_j = \sum_{j=1}^{n} f_{ij}$.

(6) Normalization to E_j use E_{max},and get the entropy which can show the importance of evaluation index j.

$$e(f_j) = {}^{E_j}/_{\ln n} = -({}^{1}/_{\ln n})\sum_{i=1}^{n}({}^{f_{ij}}/_{f_j})\ln({}^{f_{ij}}/_{f_j}) \tag{8}$$

(7) Determine the evaluation weight of evaluation index use $e(f_j)$,

$$\vartheta_j = {}^{[1-e(f_j)]}/_{(n-E_e)} \tag{9}$$

Suppose $E_e = \sum_{i=1}^{n} e(f_j)$,

and ϑ_j satisfied with $0 \leq \vartheta_j \leq 1,\sum_{j=1}^{m}\vartheta_j = 1$

3.3 Delphi Method

Delphi method, also known as an expert method, its characteristic is centralize expert experience and advice to determine the index weights, and obtained satisfactory results through continuous feedback and modify. Basic steps are as follows:

(1) Select expert. Select the exports of both practical work experience and deeper theoretical training in power grid area, there are about 15-30.Give n index which need to set weight ,related information as well as the rules of determine weight to the selected experts , asking them to give the weight of each index independently.

(2) Back the results, then calculate the mean and standard deviation of each index weight.

(3) The calculated results and the additional information returned to the experts, asked all the experts to determine the weights again on a new basis.

(4) Repeat the second and third step ,until the deviation of the weight and the mean does not exceed preset standards ,that is to say the expert's opinion basically consistent, at this time take the mean weight of each index as the index weights.

3.4 Combination Weighting Method

Suppose there are m index, the weight which obtained by the above three methods is expressed as $w_j(k)$ the combination method of calculating weight is:

$$\theta_j = \prod_{k=1}^{q} w_j(k) / \sum_{j=1}^{m}\prod_{k=1}^{q} w_j(k) \quad , \quad j = 1,2,\ldots m$$

Suppose the combined weights is θ_j, $k = 1,2,3$.

4 The Realization of Weight Calculation

4.1 Programming Framework

Weight calculation Application design for the investment evaluation and assessment of the power grid, its interface flexibility, need to fully consider the use of high efficiency, the application can be well integrated into evaluation model of power grid enterprise. The main function of this application is to calculate weight for index which obtained by normalization processing, realized Coefficient of variation method, Delphi method, Entropy method and Combined Weight Method, every method should provide external interfaces for other programs.

Application is divided into four layers in accordance with the J2EE architecture:

(1) The Client Layer

The client layer contact with user directly ,provide operation and management interface to user, allow user chose and maintain various algorithm expediently.

(2) The Web layer

The Web layer can do simple processing to the user's input , it is responsible for calling the EJB functional components of business layer, transfer or receive the information from the business layer and interact with the customer layer.

(3) The Business layer

The Business layer is the core function of the weight calculation, we should make the logic function independent while achieving the calculation, reduce coupling, and take into account the efficiency and safety. Each business component exchange information with the system through interfaces, all components communicate through the interface to ensure their independence and stability. The main business components are: the Coefficient of variation method business components, Delphi method business components, Entropy method business components, Combined weight calculation business component and Database business component.

(4) Information system layer

The Information system layer is responsible for application and database server operations. Including establish a database connection, maintain the database connection, operation the data, share resources. The business layer calling the database object components for data persistence when it needs to access data.

The framework diagram of program is shown in fig.3.

The business layer is achieved with component technology, component technology is popular with software developers in favor. EJB technology has many advantages compared with other component technology, such as good cross-platform, network communication quality, high security, high support and so on, can be completed well the task of communication with Web layer and information system layer[4], its works is shown in fig.2.

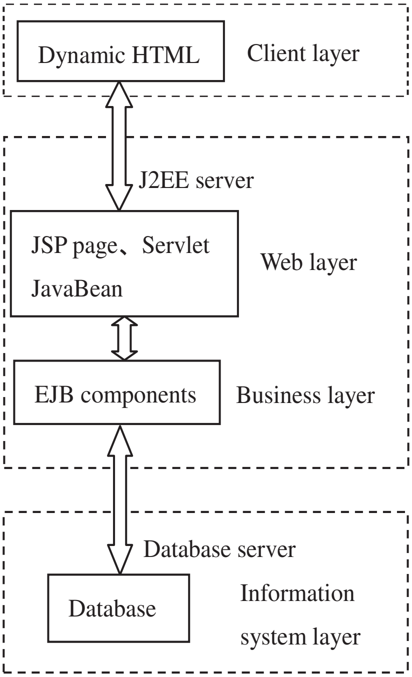


Fig. 1. The framework diagram

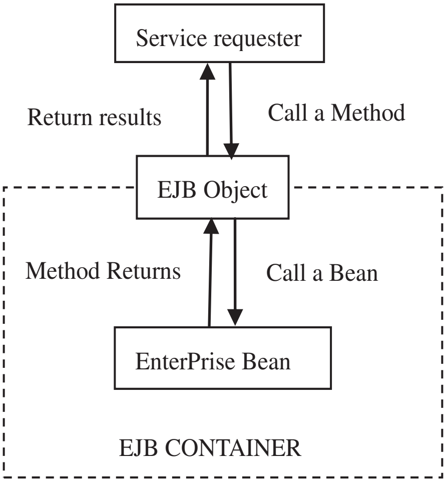


Fig. 2. EJB component workflow

4.2 Function Implementation

Component technology has the advantage of better reusability, and provide the standard application interface, so the design of the application function should be thinning as much as possible, make sure a separate component can complete certain functions independently. Component provides lots of remote interfaces, the interface define business method which would provide to the client, these methods will be implemented in Bean's implementation class, the container will generate implementation class of the interface during deployment .

Business layer component implementation details are as follows:

(1) The Coefficient of variation method business components

In the algorithm, the main function of the component is process the data matrix which is formed by the m index and n power grid evaluation object transmitted from the client, calculate Standard deviation, coefficient of variation, weight and return the results.

Remote interface are as follows:

```
@Remote
public  interface  CoefficientOfVariationRemote {
      public double[]
     CoefficientOfVariationCalculator(double[][] X);
}
```

Calculation process is shown in fig.3.

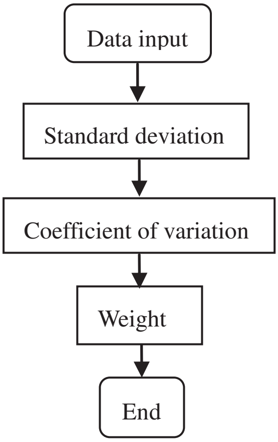


Fig. 3. Coefficient of variation method flowchart

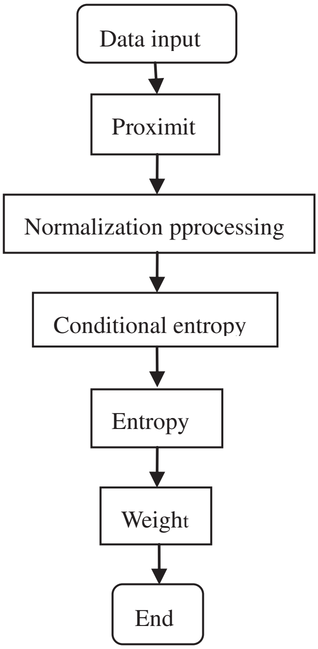


Fig. 4. Entropy method flowchart

(2) The Entropy Method business component

Concrete realization of the component is selecting the best value from the preset index value, calculated proximity, normalization processing , calculated conditional entropy, calculated entropy, finally get the index weight and returns the result. Remote interface are as follows:

```
@Remote
public interface EntropyWeigh{

public double[] EntropyWeighCalculator(double[][][] X);
}
```

Calculation process is shown in fig.4.

(3) The Delphi Method business component

Concrete realization of the component is calculated index weights developed by experts ,and get the mean and standard deviation.

Remote interface are as follows:

```
@Remote
public interface DelphiMethod{
public double[][] FirstGroup(double[][] X);
public double[] SecondGroup(double[][] X);
}
```

Calculation process is shown in fig.5.

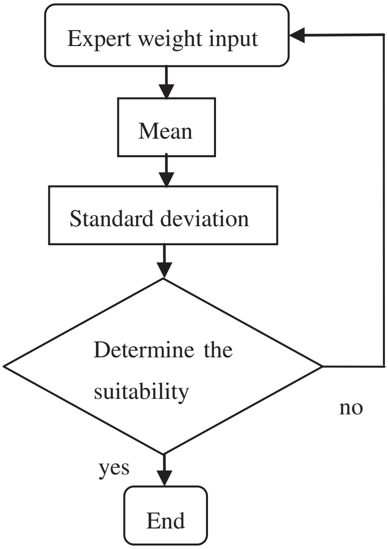


Fig. 5. Delphi Method Flowchart

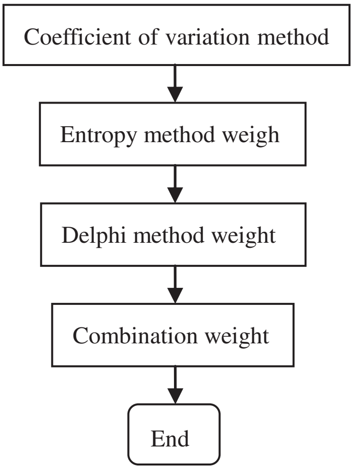


Fig. 6. Combination weight calculation flowchart

(4) Combination weight calculation component

Get the weight calculated from above three methods, calculated Combination weights.

Remote interface are as follows:

```
@Remote
public interface CombinationWeighting{
public double[]
       CombinationWeightingCalculator(double[][] X);
}
```

Calculation process is shown in fig.6.

5 Conclusion

Index Weight Calculation Model is based on mathematical theory, Computer calculation of complex weights can help reduce the workload of manual, improve the accuracy of calculation ,avoid human error resulting from the calculation. The program provide service to the Power Grid enterprise evaluation model, simple operation, flexible interface design, good usability[5].

In the project evaluation process, it is important to determine reasonably the index weight. As a measure of the performance and monitoring objectives, Evaluation is bound to involve a number of specific objective micro-elements. The role of these factors to evaluation the object is direct and hidden, and not for the transfer of consciousness. Therefore, we should pay more attention to how to calculate a more accurate evaluation index weights, or can not be avoided some of the evaluation index[6]. The three algorithms used in this article, the final combination of the three algorithms calculated, basically we can think it is a very objective way.

References

1. Liang, L., Ruiming, W.: A study on Eliminating Correlation of Evaluation Index and Obtaining Modified Weights. Shanghai Jiaotong University (2009)
2. Ming, X., Hongwei, Z., Xiaofeng, W.: The Research of Entropy method to determine weight. Market modernization (2007)
3. Yinguo, L., Xinchun, L.: The Reserch of Weight Determination Method in Comprehensive Evaluation. The liaodong college journals (2007)
4. Huoming, L.: The classical of introduction to EJB 3.0, pp. 10–146. Tsinghua university press, Beijing (2008)
5. Mingxia, Z., Wei, S., Zengqiang, M.: Component Technology Applications in Power Systems. Power Information, 63–65 (2001)
6. Lu, G., Jianming, W., Xueyan, Z., Cheng, L.: The Research of Balance Process and Auto-Implementation of Evaluation Index Weight Value Adjusting. Oranance Engineering College (2006)

An Improved Algorithm for Session Identification on Web Log

Yuankang Fang[1,2] and Zhiqiu Huang[1]

[1] College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, 210016, Nanjing, P.R. China
[2] Department of Mathematics and Computer Science, Chi Zhou University, 247000, Chizhou, P.R. China
Fyk80@163.com

Abstract. As regards session identification method on web mining, an improved one has been put forward. Firstly, considering website structure and its content, page access time threshold will be reached after collecting access time of each page, which should be used to divide sessions into various sets. Then, the session sets will be optimized further, with the help of session reconstruction, namely union and rupture. It has been proved through experiment that the session set which is attained by the above method is more faithful.

Keywords: web mining, data preprocessing, threshold, sessions.

1 Introduction

Internet is an open, dynamic and heterogeneous network worldwide where information distribution is dispersive, while there is not a unified management organization, which results in difficulties to access helpful information. Most people are high likely to be lost in it, and might be bored with the jumping access or impatient to wait for finding his/her information. One effective way to solve those problems is web mining which means that data mining method be applied to web data so that useful model and hidden information will be gained. Web mining may be classified into the following three kinds: Web content mining, Web structure mining and Web usage mining. The third one means mining web log record to find out web-user's access model, through which designer's knowledge, the degree of user's interest and their access way might be found out, and some decision-making information, being useful to web-designers and dealers, will be gained, which is on how to optimize web structure, carry out personalized service and access control. While mining web log record, data preprocessing is the most important and most time-consuming part, and the faithfulness on session identification in it is the key element influencing the result of mining web log.

2 Traditional Session Identification

Session refers to a series links made by one web-user to get information on certain topic[2],[3].Under the condition that there is a longer time spam in the web log which

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 53–60, 2010.

users might access for many times, the task of session identification is to identify the accessing with the same aim of the same user. At present, there are the following 4 arithmetic:

2.1 Hvist

User’s access time to the whole website will be given an upper limit---θ. If it is exceeded, it will be thought as a new session[3],[4]. Suppose the timestamp of the beginning page in a session is t0 and the same user’s URL requirement time is t, it will be in the present session under the condition of the inequality, namely $t-t0\leq\theta$, while the first page which is in line with this inequality $t0+\theta<t$, it will be the beginning page of the following session. Usually θ is 30min.

2.2 Hpage

User’s access time to one page will be given an upper limit---△t[5]. If the interval between two successive accessing does not go beyond the △t, they belong to the same session; Or they belong to two sessions. Usually △t is 10min.

2.3 HRef

It will be classified according to user’s access history and reference pages.[3] If one accessing of user’s fail through the links on reference pages, it is high likely to belong to another session. In other words, it is the beginning of another session for the accessing to the reference page at that time is not among those which have been accessed before.

2.4 MF (maximal forward references)

The page which has been accessed before will not be among the user’s session. If the user move forward to access the next page and click the button of BACKWARD, it means that session is over and a new one will begin.

The shortcoming of the above session identification methods lies in the following two aspects: the first one is that some recordings which should belong to one session might be classified into various sessions; the second is that some recordings which should not be put in one session might be classified into one session. If there are many unfaithful parts in the session set reached through the above method, the theory value of the mined results will not be high, even it will be lost. Therefore, a new session identification method, based on access time threshold and session reconstruction, is put forward. With experiment as proof, session sets reached through the method are more faithful.

3 Session Identification Method Based on Access Time Threshold and Session Reconstruction

The method is made up by two steps. Firstly, session sets are reached through the method which is based on page access time threshold; secondly, new, more faithful session sets will appear as the result of reconstruction, namely union and rupture, of the above session sets.

3.1 Primary Session Sets Are Reached through the Method Which Is Based on Page Access Time Threshold

According to collected page access time and the importance of pages which is determined by page content and website structure, time threshold is set to each page, which is the base of identifying session.

3.1.1 Page Stay Time (T) Is Gained by the Method of Collection

The performance simulation experiment of this paper is carried out with the access recording of 10 days from October 23th ,2009 to November 3rd of the web log of the Chizhou College website ---- http://211.86.192.12 as its data. The page access time is distributed normally which has been proved through hypothesis and checkout.

According to the statistics, the data value t which can cover the 94% of the sample set is chosen as the reference value of δ the page access time threshold. The t multiplied by the certain smooth coefficient α makes δ . Choosing 94% as the covering rate is to remove the maximal value point in the data range t, while the certain smooth coefficient is a compromise to make up for some reasonable data which might be deleted. In experiment, the smooth coefficient is chosen from 1.0 to 1.5; and it has been proved through experiment relatively reasonable to choose 1.25 as α .

3.1.2 The Page Threshold δ Is Gained through the Combination of Page Content and the Structure of the Website

Until now, the δ is determined only by the collected recordings of users' access. The influence of the page importance and the website structure has not been taken into consideration.

Definition 1: linking content ratio（RLCR） refers to ratio of page content to the amount of page linking. While the size of page is SDS, the calculation formula of RLCR is as follows:

$$R_{LCR}=（L_I+L_O）/S_{DS} \quad (1)$$

Here the amount of link-in which is recorded as LI refers to the amount of pages linked to some page; the amount of link-out which is recorded as LI means the amount of linking included by some page.

If the size of one page, SDS, is 3Kbyte, its LI 2 and its LO 4, the result of RLCR will be 2. Usually, the link-in is more important than the link-out, so a weighting adjustment on them is needed. In this paper, the hypothesis on the weighting ratio of link-in and link-out is golden mean. The calculation formula of RLCR is adjusted to the following:

$$R_{LCR}=2（0.618L_I+0.382L_O）/S_{DS} \quad (2)$$

In order to apply it to the adjustment of the threshold δ, the value of RLCR is needed to be mapped in (0,1) with many mapping means available. For example, the ratio of RLCR value to the maximal among all RLCR value can be mapped in (0,1). However, it tends to be influenced by some isolated point. When the RLCR value of some page is very big, it will influence other points. The following mapping method is applied in this paper, with the influence factor of RLCR to δ being β .

Definition 2 : β is the influence factor of page RLCR to access time threshold δ, with its calculation formula[7] as β=1-exp(-RLCR). But this change to numbers while RLCR>20, β is always 1. In other words, all numbers over 20 will be 1 while they are mapped in (0,1), which means it is impossible to distinguish them in the section of (0,1). Considering it, the formula is updated to the formula(3) so that the accuracy of mapping will be improved much as the upper limit to distinguish RLCR can reach nearly 500.

$$\beta = 1\text{-exp(-sqrt(sqrt(}R_{LCR}\text{)))} \quad (3)$$

Taking the above adjustment into account, the calculation formula is as follows:

$$\delta = \alpha\, t(1+\beta). \quad (4)$$

3.1.3 Users' Session Set Is Gained on the Base of Page Access Time Threshold δ

To set the access time threshold δ of every page, collected page access time t must be got first with its set recorded as St={t1,t2,…,tn} , while δ should be adjusted under the consideration on β ,the influence factor of page RLCR, which is recorded as S_β ={ β1, β2, …, βn} . The calculation steps are as follows:

Firstly, the set of t, St, is gained on the base of the collection of web log which is classified according to users, namely user identification. With the method in this part, St is gained after collecting page access time;

Secondly, page access time threshold set, S δ ,is gained after adjusting St on the base of influence factor set S_β and α , which is done according to formula (4);

Finally, users' session set is gained on the base of web log which is classified again according to S δ

3.2 Reconstruct Session to Improve Primary Session Sets Ses

3.2.1 Unite

During the process of session identification, recordings in one session is might be divided into two different sessions. For example, one actual session, <L1, …,Li,Lj, …,Ln>,in which L1,L2. …,Ln are recordings, might be divided as two sessions, <L1, …,Li > and <Lj, …,Ln>.

Li and Lj are in one actual session, which means the user has not turn to another topic, or he/she has not left the website. Simply speaking, there exists direct or indirect linkage between Li and Lj in the topological structure of the website.

Considering the above facts, while improving sessions, if meeting session borderline Li and Lj,(they are expressed as < …,Li > and <Lj, …>, it should take two kinds of situation into account to put Li and Lj into one session.

- If the mode of users' regular access is Li-> Lj, which is the frequent access mode, Li and Lj should be united directly; or the second judge should be carried out. The mining of frequent access mode attracts great attention. At present, a better arithmetic of time complexity degree is Suffix Tree mining the Maximal Frequency Sequence, in short MFS, brought forward by Xiao Yongqiao, with time complex degree being 0(nlogn). The mining method will be applied to sessions formed in 3.1 part to get MFS file; then the borderline of neighboring sessions will be decided if it is frequent access mode.
- With the consideration on topological structure of website, make the judgment that if recordings of Li or before it can link to Lj, which is called as whether tracing to Lj, indicated as function Trace(Li,->Lj,L) in arithmetic. If the above recordings are found, Li and Lj will be united; or the original classification will be defaulted. As to the value of L, it will be explained in the part of experiment analysis.

3.2.2 Rupture

In the process of session identification, it is also possible that two or more requesting in various sessions is classified into one same session. For example, there are two actual sessions <L1, …,Li > and <Lj, …,Ln>, L1,L2…,Ln being recordings, which are classified into one session, namely <L1, …,Li,Lj, …,Ln>. Although Li and Lj are two sequent recordings of one user on the same server, user has turned to another topic for they are not in the same actual session. In the way of certain amount of drawing back, or entering the website and etc., the user has reached the page recording Lj.

Considering the above facts, as regards Li and Lj, shown as < …,Li ,Lj, …>, the internal recordings of session, if L recordings passing Li or before Li can not be traced to Lj, Li and Lj will be ruptured.

Besides, if users turn from one topic to other topics, usually it will last for a long time t. In other words, the interval between Li and Lj is t. So it is not necessary to make the judgment on all two enamoring recordings in session, while that recording between which the interval is relatively long is needed. Here, the value of T is the access time value of present webpage.

In conclusion, it must make the obtained session be close to actual session to improve sessions through the two methods, union and rupture.

3.2.3 Session Reconstruction Arithmetic

The improved arithmetic of session identification will be given in the following. In the event the two neighboring sessions take the form of < …,Li > and <Lj, …>, it is time to make a decision if it is needed to unite Li and Lj into the same session. If one among the following two conditions is met, union will be carried out: (1)the mode Li ->Lj is frequent. In other words, the mode may be found in MFS; (2)Lj may be traced back through the recording Li or recordings before Li.

Three properties of every page will be taken into account, including url_id which means URL page serial number of recording on page access, time which refers to page access time threshold, and status, the integer domain showing the position of the recording in session, which is 1 when it is at the beginning of the session, 2 at the end of it and 0 otherwise.

The arithmetic may be described as follows:

Algorithm Session Restructure

```
Input: the session set Ses appeared in 3.1 part, access time threshold set S$\delta$ of every web page,
topology structure file of the website and the maximal frequent access serial file MFS.
Output: the improved session set Ses.
1  readitem(Ses,R1);i=0;
2  do while (Ses should not be empity)
3  i=i+1
4  readitem(Ses,R2)
5  ti=R2. time-R1. time; $\delta$ i=readitem(S $\delta$ , $\delta$ i);
   // the i in $\delta$ i is the same with the one in URL_id.
6     if (R1.status=2 and R2.status=1)
          // it will be found out that if R2 and
          //R1  are the beginning or the end of the two neighboring sessions.
7     if  (R1.url->R2.url in MFS) or (Trace(R1.url->R2.url,  L)
8       R1.status=R2.status=0   //union
9     endif
10  else  //R1 and R2 belong to the same session.
11    if  ti>= $\delta$ i  and  (!Trace(R1.url->R2.url,L))
12      R1.status=2;R2.status=1;  //rupture
13    endif
14  endif
15   R1=R2
16  enddo
```

The arithmetic will open the session set Ses, website topology structure file and the mined maximal frequent access serial file MFS first, and read the 1st recording; the 4th line is to read other recording of the user's session; the 5th line is to calculate the actual page access time of the number i recording and read the page access time threshold of the recording; the 6th line is to judge if the two recordings are the borderline of the two neighboring sessions; the 7th line is to decide if they meet the condition of union; the 8th line is to change the domain value of status to unite R1 and R2; the 11th line is to find out that if it meet the requirement of rupture; the 12th line is to make the rupture; the 15th line is to let record R2 be the value of R1 and return to the 2nd line, which should be done in circle.

The role of all functions in the arithmetic will be explanation in the following:

Readitem(Ses,R1) is to read recordings in session set and store them in R.

Trace(R1.url->R2.url, L) is to find the linkage of R1 and R2 in topology structure of website. If it is not found, the linkage between R2 and No.L recording before R1 will be sought. In this way, it will not stop until the number L-1(L is an integer which is not less than 1) recording before R1 is found. In the event of finding the linkage, the function returning value is true, or it is false. Where there are various values of L, there will be different results of sessions. Parameter L may be fixed on actual situation.

According to experiment, if the topology structure of website and actual access amount are taken into account, it is better to make L 1,2 or 3, while it is the best to take L=2 in this experiment.

4 Experiment Result and Analysis

The data of this experiment comes from Chizhou College website: 211.86.192.12. The server log data is from October 24, 2007 to November 3rd.In the experiment, it makes a comparison on the result of four session identification methods simultaneously, namely session identification arithmetic based on citation (method 1), session identification arithmetic based on fixed time threshold ten minutes(method 2), session identification arithmetic based on page access time threshold δ (method 3), session identification arithmetic based on page access time threshold and session reconstruction (method 4).

The present popular evaluation standard[4] is applied in this paper: the degree of session complete reconstructed by arithmetic h. Usually, two indexes, precision degree and recall ratio, are taken to measure the degree of reconstruction[8]. Precision is the ratio of number of complete constructed sessions to the number of total sessions got through construction: precision(h)=|Rh∩R|/|Rh|. recall ratio is the ratio of number of complete constructed sessions to the number of real sessions: recall(h)=|Rh∩R|/|R|. Data of the experiment is shown in table 1, while making comparison of all arithmetic takes the method based on citation as benchmark. The experiment shows that method 4 is better than other methods, no matter from the aspects of precision or recall ratio.

Table 1. Outcome Of Comparison On All Methods Of Session Identification

Methods of session construction	amount of sessions	amount of session intersection	precision%	recall ratio%
based on citation(R)	|R|=26953	|R∩R|=26953	|R∩R|/|R| =100	|R∩R|/|R| =100
based on fixed time threshold(T)	|T|=43850	|T∩R|=11286	|T∩R|/|T| =25.738	|T∩R|/|R| =41.873
based on page access time threshold(A)		A|=58874	|A∩R|=20134	|A∩R|/|A| =34.985	|A∩R|/|R| =74.7004
based on page access time threshold and session reconstruction	|FA|=59168	|FA∩R|=23043	|FA∩R|/|FA| =38.945	|FA∩R|/|R| =85.493

5 Conclusion

The paper puts forward the session identification method based on page access time threshold and session reconstruction. Firstly, based on page access time threshold, the page threshold δ is got through the combination of page content and website structure. Then users' session set is obtained on the base of the time threshold δ. Finally, the session set is improved further with the method of session reconstruction so that the session set is closer to the actual session.

Acknowledgment

The authors would like to thank the anonymous referees for the useful suggestions for improving this paper. This project is supported by the National High-Tech Research and Development Plan of China under Grant No. 2009AA010307.

References

[1] Jia-wei, H., Xiao-feng, M., Jing, A.: Research on web mining. Journal of computer research & development 38(4), 405–414 (2001)

[2] Facca, M., Lanzi, P.L.: Mining Interstiong Knowledge from Weblogs: A Survey. Data and Knowledge Engineering 53(3), 225–241 (2005)

[3] Cooley, R., Mobasher, B., Srivastava, J.: Data Preparation for Min2ing World Wide Web Browsing Patterns. Knowledge and Information system 1(1), 5–32 (1999)

[4] Fu, Y., Sandhu, K., Shih, M.: A generalization - Based Approachto Clustering of Web Usage Session. In: Masand, B., Spiliopoulou, M. (eds.) WebKDD 1999. LNCS (LNAI), vol. 1836, pp. 21–28. Springer, Heidelberg (2000)

[5] Spiliopoulou, M., Mobasher, B., Berendt, B., et al.: A Framework for the Evaluation of Session Reconstruction Heuristics in WebUsage Analysis. INFORMS Journal of Computing 15(2), 171–179 (2004)

[6] Chen, M.S., Park, J.S., Yu, P.S.: Data Mining for Path Traversal Patterns in a Web Environment. In: Proc 16th Int ’l Conf. Distributed Computing System (ICDCS 1996), pp. 385–392. IEEE CS Press, Los Alamitos (1996)

[7] Xianliang, Y., Wei, Z.: An Improved Session Identification Method in Web Mining. Huazhong University of Science and Technology Journal (natural science edition) 7, 33–35 (2006)

[8] Fang, Y., Lijuan, W.: Practical Data Mining. Electronic Industry Press, Beijing (2004)

A Chinese Web Page Automatic Classification System

Rongyou Huang[1] and Xinjian Zhao[2]

[1] College of Computer Science and Technology, Zhejian University of Technology,
Hangzhou, Zhejian, 310023, China
huangrongyou1@126.com
[2] College of Information Engineering, Zhejian University of Technology,
Hangzhou, Zhejian, 310023, China
zxj@zjut.edu.cn

Abstract. In recent years, with the popularization of development of the network, people are getting closer and closer with the net and the number of web page is increasing rapidly. To help people to quickly locate user-interesting web page promptly in the flood of web information and improve the precision of search engine, a system of Simple Bayesian classifier for automatic classification of Chinese web page is proposed. Experimental results show that the system have high page detection rate and have ability to self-learning.

Keywords: Bayesian;TF-IDF; Self-learning; Web Page Classification.

1 Introduction

As of 2009, the number of Chinese web pages has reached 33.6 billion. Facing with massive information, the way to search web information manually is not only inefficient but also spend a lot of effort but result is not satisfactory. To help users to quickly locate interesting page, a system of Chinese web page automatic classification is proposed. Research shows web page classification is one of an effective approach to use mass Internet information.

2 Chinese Web Page Automatic Classification System Overview

This system includes two modules: page pre-processing module and the classification module. In my system, all experiment pages are divided into two categories: the training set and the test set. Learning through the training set, the system get the original bayesian classifier at first. And then it uses the classifier to classify the test set and simultaneously improves classification strategies by self-learning.
The main steps of the system are as follows:

- Preparation phase: the training set of pages are manually divided into k categories according to content relevance;
- Training process: each page from the training set is processed as follows: page pretreatment, feature extraction and features saving in xml format. Then, the Bayesian classifier learns classification strategies base on them.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 61–66, 2010.

- Classification process: each page from the test set is processed as follows: page pretreatment, feature extraction and identifying category which the page belong to by Naive Bayes classifier.
- Self-learning: the test page with certain requirements is added to the training set.Then classifier updates classification rules through a new round of learning .

2.1 Vector Space Model(VSM)

VSM is one of widely used model. In this model, document *d* is expressed as n-dimensional vector (t_1, t_2, …,t_i,..., t_n). t_i is the *i*th feature in the document *d*. In the system, we use TF-IDF to calculate the weight of term[3][4]. The formula shows as follows:

$$w(t_i) = tf(t_i) * idf(t_i) = tf(t_i) * \log\left(\frac{\mathrm{N}}{\mathrm{n}} + 1\right) \tag{1}$$

In the above formula,w (t_i) is the weight of the feature t_i in the document *d*. $tf(t_i)$ is the number of the feature t_i occurrencing in the document *d*. $idf(t_i)$ is the importance of feature t_i among all the documents. *N* is the total number of all documents. *n* is the number of documents which contains feature t_i among all the document.

2.2 Bayesian Formula

Random events A_1, A_2,...,A_j,…,A_n is a division of finite sample space *S*. Then the Bayesian formula[3] as follows:

$$P(A_k \mid B) = \frac{P(A_k)P(B \mid A_k)}{\sum_{j=1}^{n} P(A_j)P(B \mid A_j)} \tag{2}$$

In the above formula,$P(A_k|B)$ is posterior probability. $P(A_k)$ is prior probability.

3 Page Pre-processing Module

3.1 Page Pretreatment

The core of page pretreatment is that we shall remove the information which doesn't make sense to content and extract useful information. Web page is a kind of text file of semi-structured form and typically contains <head> part and <body> part. In the processing, we remove the useless labels at first,for example <style>, <script>, <link> and so on. Second, we must keep <meta> and <title> label, because this information is a summary description of entire page and closely related with the content of the page. Third, <body> label not only contains body of the page, but also often is mixed with the navigation bar,advertising and other noise. Junfeng Duan’s experiments show that in a content-based (non-navigation-based) of the page, the link text is navigation or ads in most cases, and just by removing the link text can effectively remove noise. My system is using this simple and effective method.

3.2 Chinese Word Segmentation and Stop Words

Compared to English words with spaces and punctuation as separators, Chinese are much more complex syntax. It treats the word as the basic unit. There is no clear distinction identity between words. Chinese word segmentation is the basis of Chinese web page classification. Currently there are some Chinese word segmentation system and each of them has their own characteristics. My system uses ICTCLAS which developed by Chinese Academy of Sciences as the basis for Chinese word segmentation modul.

The contribution of the low-frequency words to the text features greater than high-frequency words[1]. Such as "I", "ah" and other high-frequency words, they have no help to distinguish text. But they increase the dimension of feature space to increase the computational complexity. These words are called stop words. My system select 900 common words to form stop word list. Stop word list filters stop word and it occurs between page segmentation and feature extraction. Thus, the page *d* was translated into a Chinese phrases set *T*.

3.3 Page Feature Selection

The purpose of page feature selection is that selecting the terms which strong correlation with the content from *T* to make up features of page *d*. The number of features isn't the more the better. Too many features will increase the dimension of feature vectors and it may lead to reduce the system speed. Selecting appropriate features will improve the accuracy of the system[1].

In my system, the feature extraction based on the following assumptions: the smaller attribute weights the smaller impact on the classification results[2]. First, each weight of word from set *T* is calculated. Then, the word whose weight big than predefined threshold is selected as features of page *d*. TF-IDF and DF are both commonly used to weight calculation. TF-IDF is adopted in my system. The sencond experiment proves that TF-IDF algorithm improves more classification accuracy than DF.

In my system, all pages are divided into m-categories, referred to as $C_1,C_2,\ldots,C_k,\ldots C_m$. Modify the formula (1) to conform to our requirements.

$$w'(t_i)=tf(t_i)*idf(t_i,C_k)=tf(t_i)*\log\left(\frac{\mathrm{N(C_k)}}{\mathrm{n(t_i,C_k)}}+1\right) \tag{3}$$

In the above formula, w'(t_i) is the weight of the word t_i in the document *d*. $tf(t_i)$ is the number of the word t_i occurrences in the document d. $idf(t_i,C_k)$ is the importance of word t_i among the documents which belong to C_k. $N(C_k)$ is the total number of the documents which belong to C_k. $n(t_i,C_k)$ is the number of documents which contain word t_i among the document belong to C_k.

Research shows that label <title> and <metal> have more contribution to web features than the body of the page. Therefore, head words are given a higher weight factor. Finally,the formula of word weight calculation shows as follows:

$$W(t_i)=\begin{cases} w'(t_i)*\alpha & t_\mathrm{i}\in head \\ w'(t_i)*\beta & t_i\notin head \end{cases} \qquad (4)$$
$$1>\alpha>\beta>0;\alpha+\beta=1$$

In the above formula, $W(t_\mathrm{i})$ is the weight of word t_i in the document d. The head is the set which contains the words which get from label <title> or <metal>. My system take α: β = 9:1. In order to facilitate processing by the classifier, page features are saved as xml format.

4 Classification Module

4.1 Naive Bayes Classifier

First, Bayesian respectively calculate posterior probability P (C$_\mathrm{k}$ | d) of page $d = (t_1, t_2, \ldots, t_\mathrm{i}, \ldots, t_\mathrm{n})$ belong to different category. Then according to Bayesian maximum posteriori criterion, we determine which category the page belong to. The Naive Bayesian formula shows as follows:

$$C_{nb}=\operatorname*{argmax}_{k=1}^{m}\{P(C_k \mid d)\} \qquad (5)$$

According to Bayesian assumptions[5], posterior probability formula is modified as follows:

$$P(C_k \mid d)\propto P(C_k)\prod_{i=1}^{n}P(t_i \mid C_k)=\frac{N(C_k)}{\mid N\mid}\prod_{i=1}^{n}\frac{n(t_i,C_k)+1}{\mid N(C_k)\mid} \qquad (6)$$

In the above formula, N(C$_\mathrm{k}$) is the total number of the documents which belong to C$_\mathrm{k}$. |N| is the total number of training pages. $n(t_\mathrm{i}$, C$_\mathrm{k}$) is the number of page containing feature t_i in the C$_\mathrm{k}$.

Compared to other classification methods, Naive Bayes classification algorithm does not need search, just simply calculating the frequency of occurrence of each attribute among the training examples to estimate the probability of each attribute. So Naive Bayesian classification algorithm's efficiency is particularly high.

4.2 Self-learning Strategies

During the classification processing, the so-called self-learning is to put the page which has been classified into the corresponding training category and force the classifier strategies updated by a new round of learning. However, classification results may exist error. So in order to avoid putting page into error category to cause more errors. Therefore, the system only puts the page which closely related with the falling category and obviously different with other categories into the training set.

Using previous section's formula to calculate posterior probability of page d with each category.The results is P(C$_1$|d), P(C$_2$|d),…,P(C$_\mathrm{k}$|d),…P(C$_\mathrm{m}$|d). The biggest posterior probability[6] P$_\mathrm{max}$(Cx|d) means page d belong to C$_\mathrm{x}$.

$$dif = \min_{k \neq x} \left| P_{\max}(C_x \mid d) - P(C_k \mid d) \right| \tag{7}$$

When the *dif* exceeds a certain threshold, it put the page to the corresponding category training set. Experiment one shows that classification self-learning is better.

5 Classification Module

We collect 2400 Chinese web page as our experimental data set from Internet. These pages contain 6 category and each category contains 400 pages. 100 randomly selected web pages as training set and the remaining 300 pages as a test set. Experiments use classification accuracy to estimate the classification result:

classification accuracy = correct classified page / all test page.

Experiment one: each category randomly respectively selects 50,100,150,200,250 and 300 pages from the test set to test and respectively statistical classification accuracy. The result as follows:

Table 1. Experiment result on self-learning

	50(%)	100(%)	150(%)	200(%)	250(%)	300(%)
Automobile	90	92	88.6	89	90.2	90.3
Finance	80	84	84	84.5	85.2	87
IT	84	86	85.3	85	85.6	86.7
Health	94	90	87.3	89.5	90.8	91.7
Sports	86	82	82.6	84	84.2	84.3
Travel	88	86	86	86	87.2	87.7

As can be seen from above table, we know that along with the number increase of test pages, the classification accuracy is getting better. It means the classifier achieve self-learning during classification processing and the classification accuracy can meet practical applications.

Experiment two: first,select 300 pages as the test set. Then we respectively calculate classification accuracy in three case. The first case uses TF-IDF to gain page features and achieves learning function. The second case uses DF to gain page features and achieves learning function. The third case uses TF-IDF to gain page features but doesn't achieve learning function. The result as follows:

Table 2. Experiment two

	First(%)	Second(%)	Third(%)
Automobile	90.3	89	78.3
Finance	87	85.3	80
IT	86.7	86	82.7
Health	91.7	87.7	85.3
Sports	84.3	83.7	79
Travel	87.7	83	81

Comparison of the first and the second case, we know accuracy of classifier using TF-IDF to get feature better than DF. Comparison of the first and third situation, we also know the accuracy of classifier with self-learning better than not. Therefore, the first case using TF-IDF and self-learning is best. My system adopts the first method.

6 Conclusion

Chinese web page classification is a system of practical significance. It can be applied to the meta search engine or a single search engine. Then automatic clustering search results will be returned to the user. Category's name and a summary page is displayed clearly to the user, the user can modified search strategy to meet their own requirements.

References

1. Alina Lupascu, C., Tegolo, D., Trucco, E.: A Comparative Study on Feature Selection for Retinal Vessel Segmentation Using FABC. In: Jiang, X., Petkov, N. (eds.) Computer Analysis of Images and Patterns. LNCS, vol. 5702, pp. 655–662. Springer, Heidelberg (2009)
2. Shih, L.K., Karger, D.R.: Using urls and table layout for web classification tasks. In: Proceedings of the 13th international conference on World Wide Web, pp. 193–202 (2004)
3. Tarau, P., Mihalcea, P., Figa, E.: Semantic document engineering with WordNet and PageRank. In: ACM symposium on Applied computing, pp. 782–786 (2005)
4. Menczer, F.: Combining link and content analysis to estimate semantic similarity. In: World Wide Web conference on Alternate track papers & posters, pp. 452–453 (2004)
5. Zhang, H., Jiang, L., Su, J.: Augmenting naive Bayes for ranking. In: International conference on Machine learning, pp. 1020–1027 (2005)
6. Jian-Shuang, D., Qi, L., Hong, P.: Information retrieval from large number of Web sites. In: ICMLC 2005, pp. 2172–2177 (2005)

Towards a Quality-Oriented Real-Time Web Crawler

Jianling Sun, Hui Gao, and Xiao Yang

VLIS Lab
College of Computer Science and Technology
Zhejiang University, Hangzhou 310027, China
{sunjl,gaohui,yang_xiao}@zju.edu.cn

Abstract. Real-time search emerges as a significant amount of time-sensitive information is produced online every minute. Rather than most commercial web sites having routine content publish schedules, online users deliver their postings on web communities with high variance in both temporality and quality. In this work, we address the scheduling problem for web crawlers, with the objective of optimizing the quality of the local index (i.e. minimizing the total weighted delays of postings) with the given quantity of resources. Towards this, we utilize the posting importance evaluation mechanism and the underlying publish pattern of data source to exploit a posting weights generation prediction model, which is leveraged to help web crawler decide the retrieval points for better index quality. From extensive experiments applied on several web communities, we show the effectiveness of our policy outperforms uniform scheduling and the one purely based upon posting generation pattern.

Keywords: web data mining, posting weight, publish pattern, crawl scheduling.

1 Introduction

A problem created by the rapid pace and huge volume of information created by real-time Web technologies and practices is finding high-quality information timely. One approach, known as real-time search, is the concept of searching for and finding information online as it is produced. Among all kinds of heterogeneous data sources, web forums and web communities are two typical portals for online users to share their information and opinions in public. Every minute hundreds of thousands of postings are created by millions of users on web forums or web communities. Most of these postings are time-sensitive. For example, postings from online stock communities are associated with the current stock index and can reflect future trends. From search engine's perspective, how to effectively index such information becomes a valuable issue.

Policies for maintaining freshness of local index are traditionally either push-based or pull-based[1]. Push-based policies call for pushing data updates to search engine by content providers, which are taken by some eminent SNS websites, e.g. Twitter, Facebook, by supporting an open protocol called "pubsubhubbub" proposed by Google[2]. For those classic websites without supporting such protocol, pull-based policies are still the first choice. Pull-based policies require web crawlers to contact

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 67–76, 2010.

content providers to check for updates. How to balance the freshness of local index and re-visit frequency is always the metric for a good crawl scheduling. Rather than most commercial websites having routine content publish schedules, which could be used to collaborate with search engines to enhance the efficiency and effectiveness of the crawling schemes, the contents generated from web forums and web communities are characterized as high variance in both temporality and quality. Knowing when to check for updates is fundamentally linked to the freshness of the index, so exploiting the underlying publish pattern to describe the temporal distribution regularity is the premise for a sophisticated crawl scheduling.

In addition to time-sensitivity, relevance is still the core metric for evaluating search results. The quality of postings varies due to poor regulations, and noisy data such as off topic postings or spams are apt to be delivered over and over again on web forums or web communities. It's not proper to neglect the weight of posting itself based on the assumption that information from the same data source leads to the same level of user satisfactions. Some researches utilize PageRank algorithm to give priority scores to web pages, but it proves useless in the domain of web forums or web communities, because there are no explicit links between postings[3].

Our study exploits the posting history of multiple data sources and evaluates the importance of each posting along the trace. Finally we propose a prediction model which proves more precise describing these posting weights' temporal pattern as well as eliminates side effects caused by noisy postings, e.g. spams, ads. We apply this prediction model to schedule web crawler's retrievals and achieve a better experimental result than the other scheduling policies.

1.1 Related Work

Cho and Garcia-Molina[4][5][6] deems that a homogeneous Poisson process with a constant rate is a good model to describe the probability of web page changes. The time-independence property of homogeneous Poisson model does not hold in a real-time scenario while retrieval interval is in minor granularity. The posting rate goes through wide fluctuations depending on the time of the day and the day of the week. Our study assumes that the postings are generated by an inhomogeneous Poisson process whose rate changes along with time.

Ka Cheung Sia et al.[7] proposes a periodic inhomogeneous Poisson process to model the generation of postings at RSS feeds, and employs a delay metric to evaluate the monitoring policies. Our study extends both their prediction model and quality metric by involving a factor of posting weights.

Jie et al.[8] proposes a unified online scheduling algorithm which considers the negative impact (i.e. degradation if the page is not crawled) while incorporating content providers' multi-level collaboration. In the scope of web forums or web communities, such content providers' collaboration is hard to achieve because everyone is free to post a piece of information. Additionally, our negative impact evaluation is computed by the total delays of posting weights instead of user accesses of stale contents from the search engine perspective.

1.2 Contributions

In this work, we concentrate on the scheduling problem for real-time web crawlers, with the objective of minimizing the total delays of weighted postings between generation and retrieval.

- We exploit the underlying publish pattern of online portals which are open for Internet users to deliver their postings. Due to its varying temporality and quality, we aim to find the regular pattern through history data mining.
- We propose another quality metric to measure the optimality of crawler scheduling. Instead of the original delay metric with the assumption of all postings from the same data source having an equal contribution to index quality, we evaluate the weight of posting itself. We value minimizing the weighted delays rather than merely the temporal delays.
- We revise an existing prediction model by introducing posting weights, and we apply such model in the crawl simulation experiment for online community portals, whose effectiveness outperforms the other scheduling policies.

The remainder of this paper is organized as follows. Section 2 elaborates posting weight evaluation mechanism, quality metric and problem definition. Section 3 presents our prosed algorithm to exploit the posting weights generation pattern, and develops a prediction model to determine the optimal retrievals of web crawler. In Section 4, we apply our revised scheduling policy for extensive experiments and illustrate the result and compare with the other sophisticated policies. Finally, we conclude and discuss our future plan in Section 5.

2 Framework

For real-time search engine, its crawl scheduling is not as easy as a traditional one. To capture web pages as soon as they are produced, real-time web crawler must minimize the delay between content publish online and indexed by search engine. Furthermore, posting weight is an important factor to be considered because of postings' varying qualities. How to index high-quality postings in an efficient manner calls for a better crawl scheduling policy.

2.1 Posting Weight

The importance of one posting on web forums or web communities is not decided by the number of reverse links to it as PageRank algorithm indicates, but largely depends on three unique properties.

- Number of topic participants – The number of unique persons making comments. If one person posts multiple comments, it's counted as only one.
- Number of replies – The number of total comments made on a specific posting.
- Number of clicks – The number of total mouse clicks triggered by this posting.

Normally these three factors have different proportions on deciding at which level of importance one posting is.

$$W_{Participants} > W_{Replies} > W_{Clicks}. \tag{1}$$

Most of web forums or web communities allow anonyms to participate in topic discussions, which brings about the difficulty in calculating accurate participant numbers. This work only takes reply number and click number into consideration during the posting weight evaluation.

To calculate a weight value of the *i*th posting among totally k postings, we need to know reply number$N_{Replies}$ and click number N_{Clicks} of each posting.

$$W_{Replies\ of\ i} = \frac{log_{10} N_{Replies\ of\ i}}{log_{10} \max_{j=1,\ldots k} N_{Replies\ of\ j}}. \tag{2}$$

$$W_{Clicks\ of\ i} = \frac{log_{10}\ N_{Clicks\ of\ i}}{log_{10}\ \max_{j=1,\ldots k} N_{Clicks\ of\ j}}. \tag{3}$$

$$W_i = \alpha * W_{Replies\ of\ i} + (1 - \alpha) * W_{Clicks\ of\ i}\,. \tag{4}$$

2.2 Quality Metric

Consider a data source O that generates postings at times$t_1, \ldots, t_k$. The web crawler retrieves new postings from Oat times$\tau_1, \ldots, \tau_m$. The delays associated with the *i*th posting between generation and retrieval is defined as

$$D_i = \tau_j - t_i\,. \tag{5}$$

where$\tau_j$is the minimum value with $t_i \leq \tau_j$.

The *i*th posting also has replies and clicks, which decide its weight value together according to equation (4). The weighted delay associated with the *i*th posting is defined as

$$WD_i = W_i(\tau_j - t_i)\,. \tag{6}$$

The total weighed delays of the postings from data source O is defined as

$$WD(O) = \sum_{i=1}^{k} WD_i = \sum_{i=1}^{k} W_i(\tau_j - t_i)\,, with\ t_i \in [\tau_{j-1}, \tau_j]. \tag{7}$$

2.3 Problem Definition

The optimal crawl scheduling problem could be deduced as the total weighted delays minimization problem. We use t_i to represent the *i*th posting generation time at data source O and τ_j to refer to the time of *j*th retrieval for data source O by the web crawler. Given the above definitions, we can formalize the problem of weighted delay minimization as follows:

Given the posting generation times t_i's as well as posting weights w_i's, find the retrieval times τ_j's that minimize the total weighted delays $WD(O) = \sum_{i=1}^{k} WD_i$ under the constraint that the crawler can schedule a total of *M* retrievals.

3 Proposed Algorithm

Figure 1, 2 are all based upon a week’sstatistical data about postings from a well-known online stock community.

3.1 Posting Weights Generation Pattern

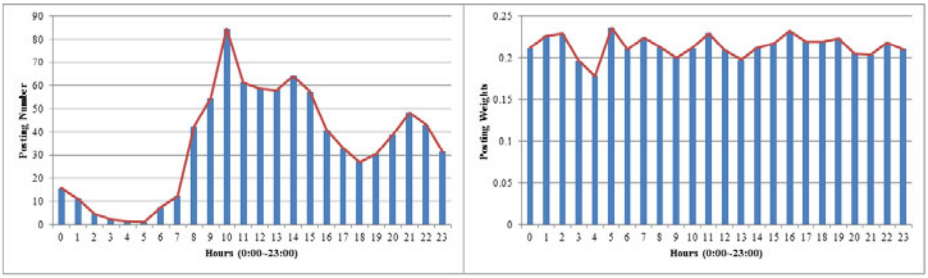


Fig. 1. The left one represents the temporal regularity of posting frequencies at different hours in a single day. During 9:00~15:00 the posting frequency is higher than the other time intervals while during 0:00~7:00 the posting frequency is the lowest. The right one shows an individual posting’s average weights at different hours of a day. We could tell that the average weight value doesn’t have positive correlations with the posting generation number of that time interval. For example, 5:00~6:00 on the timeline is associated with the smallest volume of generated postings but its average weight value is the highest. In contrast, 13:00~14:00 is characterized as more likely to generate a large number of postings but its average weight value is yet low. It infers that some postings might fall into non-peak hours but still have relatively high importance, which is implied by their significant numbers of replies and clicks. On the other hand, during posting generation peak hours there are obviously more noisy postings being produced whose weight values are minimal, such as spams.

3.2 Prediction Model

In the domain of general web search engines, significant amount of research have shown conclusively that web page changes could be modeled as a homogeneous Poisson process, whose change rate remains a time-independent constant. If the granularity is defined grosser as a month or a week, the change frequency is indeed a constant according to their theory. However, if we alter time granularity to hour, minute or even second levels, the fluctuation of posting frequency is obvious.

Upon the observation of the example in the above section, the posting weights generation pattern has certain periodicity features. We decide to adopt periodic inhomogeneous Poisson model to describe such pattern as referred to Ka Cheung Sia et al.’s work[7]. Here, we assume that the same values $\lambda(t)$ are repeated over time with a period of T (e.g. a week). That is,$\lambda(t) = \lambda(t - nT)$ for $n = 1,2, ...$This model is proved to be a good approximation when a similar pattern is repeated over time. We could use it to predict the future posting weights at a specific time point.

The value of $\lambda(t)$ largely depends on the sampling interval, and in this work we regard hourly sampling as an appropriate way. For a practical real-time web crawler the sampling interval could be adjusted much finer, e.g. minutely. To calculate posting weights generation rate$\lambda(t)$, our approach is to aggregate the historical hourly posting weights into a period coordinate (e.g. a week), and then figure out the average posting weights to obtain a graph similar to Figure 2. This discrete posting weights histogram is then approximated by a continuous function of$\lambda(t)$ through interpolation by using an ith degree polynomial function.

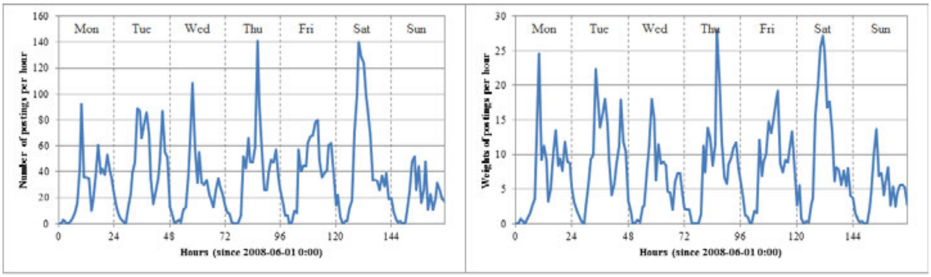


Fig. 2. The left one reveals a week's underlying publish pattern. From Monday to Saturday there is a burst of postings generated, whereas on Sunday the posting frequency is relatively low. Additionally, it roughly depicts the daily transformation of posting frequencies. Their variation forms are close to each other. Normally daytime is related with a high posting frequency while at midnight the posting rate falls down significantly. We find purely aggregating the number of postings during an interval is apt to be interfered by the varying qualities of postings. The left one is converted to the right one after postings' weights are applied. We could balance the importance and temporal regularity much better. For example, Wednesday is superior to Monday on the aspect of posting number during the peak hours as the left onesays, but from the perspective of posting weights this comparison result is just the reverse.

3.3 Crawl Scheduling

Since the web crawler has no idea about the exact times at which postings are generated and how their qualities are, it can only estimate the expected weighted delay based on the posting weights generation pattern of a specific data source. Under the inhomogeneous Poisson model, the expected weighted delay could be defined as a function of posting weights generation rate$\lambda(t)$.

For a given data source O whose posting weights generation rate is$\lambda(t)$, the expected weighted delay for the postings between the time interval $[\tau_{i-1}, \tau_i]$ is as follows,

$$EWD(O) = \int_{\tau_{i-1}}^{\tau_i} \lambda(t)(\tau_i - t)dt\,. \tag{8}$$

To derive the minimum value of$EWD(O)$, we refer to the theorem proposed by the work of Ka Cheung Sia et al.[7] to make$\frac{\partial EWD}{\partial \tau_i} = 0$for every$\tau_i$.

When scheduling M retrievals at time $\tau_1, \dots, \tau_m$ for a data source with the posting weights generation rate $\lambda(t)$ and periodicity T. When the total weighed delay is minimized, the τ_i's satisfy the following equation:

$$\lambda(\tau_i)(\tau_{i+1} - \tau_i) = \int_{\tau_{i-1}}^{\tau_i} \lambda(t)dt\,. \tag{9}$$

The above equation states the necessary conditions of an optimal solution. Based on this equation, there are two methods[7] (i.e. exhaustive search with pruning and iterative refinement) to get an optimal retrieval schedule, and in this work we adopt the first way.

4 Experiments

The objective of our experiments is to investigate the impacts of posting weights for a real-time web crawler's scheduling policy. The goal of our revised algorithm is to minimize weighted delays, which consequently decide the quality of search engine's index. In addition, we evaluate three different crawl scheduling policies targeted at different boards of a famous online stock community.

4.1 Experiment Setup

Posting Trace Our crawler has downloaded all the postings from different boards of a well-known online stock community across Jan 2008 to Dec 2009. There are rare occasions for posting deletions on these portals. Each posting is labeled with its posting timestamp as well as the number of replies and clicks.

Table 1.shows the data source information including its url and the total number of postings.

Table 1. Data Source Information

Data Source	Posting Count
http://guba.eastmoney.com/topic,000002.html	122,689
http://guba.eastmoney.com/topic,600028.html	123,498
http://guba.eastmoney.com/topic,601318.html	116,526
http://guba.eastmoney.com/topic,601857.html	156,575

Posting Weight Evaluation We develop a parser to extract the number of replies and clicks from the original postings, and calculate the value of weight according to equation (4). Here, we set $\alpha = 0.6$.

Dataset Partitions We divide these raw data into train dataset and test dataset in accordance with their timestamps. We use all the postings produced in the year of 2008 as train set and the remaining as test set.

4.2 Posting Weights Distribution

In order to better analyze the impacts introduced by the posting weights, we display both the posting distributions along different weights as well as their temporal distributions in Figure 3.

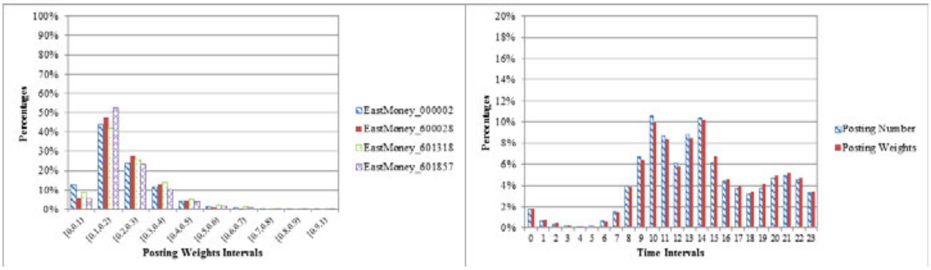


Fig. 3. The leftone describes the percentage of posting numbers along different weights. The different texture of the columns represents an individual data source. On average, 46.75% of postings converge at the weight interval between 0.1 and 0.2. Only 3% of postings are above the weight value of 0.5. From this observation we could conclude that though there are a mass of postings produced online every day, only a few postings are genuinely valuable.The right one reveals the temporal distribution characteristics of both posting number and posting weights. Over 50% of postings and posting weights are generated between 9:00~15:00. The posting weights distribution is almost the same as the posting number. A careful study finds out that during the peak hours, i.e. 9:00~15:00, the percentage value of posting weights is slightly beneath the corresponding posting number. This indicates among a convergence of postings usually exist low-quality ones which impact the total weights associated with that time interval. On the other hand, during non-peak hours for posting generations, i.e. 18:00~23:00, the percentage value of posting weights is slightly above the corresponding posting number. That means high-quality postings contribute more to such time interval.

4.3 Prediction Model Train

We use the train dataset to derive the posting generation and posting weights generation prediction models for each data source. We build weekly posting number and posting weights histograms at a granularity of one hour by history data statistics. Then we work out continuous posting generation function $\lambda(t)$ and posting weights generation function$\lambda'(t)$by polynomial approximation.

We get a set of optimal retrieval points according to equation (9) by exhaustive search with pruning. All possible plans are evaluated by exhaustively trying all choices for the first two retrieval points against our train dataset. Finally we get a crawl scheduling that leads to the minimum weighted delays.

4.4 Comparison of Scheduling Policy

To investigate further how much improvements we could get from our optimization by involving posting weights as proposing a prediction model, we now compare the average weighted delay of the following three policies:

- Uniform scheduling: The retrieval points are scheduled at uniform intervals. This policy could be considered as a baseline.
- Posting generation guided scheduling: The retrieval points are optimized based on data sources' posting generation pattern without considering individual posting's importance.

- Posting weights generation guided scheduling: The retrieval points are optimized based on data sources' posting weights generation pattern.

To employ these three scheduling policies, we suppose a fixed number of retrieval times are given during a period of time. With the determined resources allocated, we compare the final weighted delays introduced by three different strategies of scheduling to make a conclusion which one is the best.

We choose our test dataset to evaluate the effectiveness, because all the postings in the test suite have objective importance evaluation with stable click and reply numbers, so that no human evaluation is necessary in our experiment. This is only a crawl simulation rather than a real-life crawling for the sake of undetermined importance value of those just emitted postings. In that case we derive the following results based on the above three scheduling policies as Figure 4 shows.

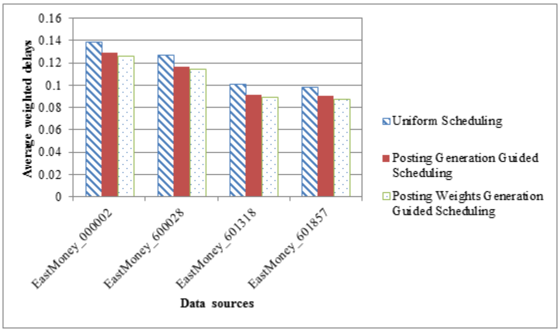


Fig. 4. This figure shows the final results which measure the average weighted delay between publish and retrieval which is achieved by different crawl scheduling policies for each data source

A comprehensive set of evidence shows that our proposed crawl scheduling policy outperforms the other two. For uniform scheduling, the scheduler dispatches the crawler at uniform intervals, which means it treats all the time points equally. When the crawler is dispatched at some non-peak hours for posting weights generation, it contributes little to the index quality, for not only very few postings are collected but also these postings might be of low quality. For posting generation prediction model guided scheduling, the average weighted delay drops significantly compared to the baseline, due to its retrieval points converge at peak hours while stay sparse at non-peak hours. It cares little about the quality distribution of posting weights across the timeline, so that some low-quality postings emitted at peak hours are apt to be crawled as important ones. For posting weights generation prediction model guided scheduling, the average weighted delay is slightly better. Profit from a better prediction model which describes the possibility of when important postings are generated, it achieves smaller weighed delays.

5 Conclusion and Future Work

In this paper, we address the scheduling problem for web crawlers, with the objective of optimizing the quality of the local index (i.e. minimizing the total weighted delays of postings) with the given resources. On the basis of posting generation pattern, we introduce a posting weight evaluation mechanism to appraise each posting's importance and study its impacts to local index quality. Based on this we build a prediction model capable of describing the trend of posting weights generations, and apply this model in web crawler scheduler and achieve a better average weighted delay value over the other two competitive scheduling policies.

All our experiments are upon the real posting traces we collected from multiple boards of a well-known online stock community. In the near future, we plan to incorporate more real posting traces in different domains, e.g. news communities. We will also extend the importance evaluation algorithm to assign a more precise weight to each posting.Furthermore, we attempt to cluster some time slots along the timeline with a similar posting weights generation pattern through machine learning, which could become a good reference for crawl scheduling.

References

1. Bright, L., Gal, A., Raschid, L.: Adaptive Pull-Based Data Freshness Policies for Diverse Update Patterns.Technical Report, UMIACSTR-2004-01, University of Maryland
2. PubSubHubbub protocol, `http://code.google.com/p/pubsubhubbub/`
3. Chen, Z., Zhang, L., Wang, W.: PostingRank: Bringing Order to Web Forum Postings. In: Li, H., Liu, T., Ma, W.-Y., Sakai, T., Wong, K.-F., Zhou, G. (eds.) AIRS 2008. LNCS, vol. 4993, pp. 377–384. Springer, Heidelberg (2008)
4. Cho, J., Garcia-Molina, H.: Synchronizing a database to Improve Freshness. In: SIDMOD Conference (2000)
5. Cho, J., Garcia-Molina, H.: Effective Page Refresh Policies for Web Crawlers. ACM TODS 28(4) (2003)
6. Cho, J., Garcia-Molina, H.: Estimating Frequency of Change. ACM Transactions on Internet Technology 3(3) (2003)
7. Sia, K.C., Cho, J., Cho, H.-K.: Efficient Monitoring Algorithm for Fast News Alerts. IEEE Transactions on Knowledge and Data Engineering 19(7) (2007)
8. Xu, J., Li, Q., Qu, H., Labrinidis, A.: Towards a Content-Proivder-Friendly Web Page Crawler. In: Proceedings of the Tenth International ACM Workshop on the Web and Database (2007)

Research on the Value of Business Information Online Based on Fuzzy Comprehensive Evaluation

Xiaohong Wang, Yilei Pei, and Liwei Li

College of Management, Beijing Union University, Beijing, China
gltxiaohong@buu.edu.cn, gltyilei@buu.edu.cn,
gltliwei@buu.edu.cn

Abstract. With the rapid development of network and e-commerce, the value of business information online has already attracted attention of small and medium-sized enterprises, because it is the basis of effective use of information and is the basis of decision making for managers to scientifically evaluate the value of business information online. This paper provides an evaluation index system of the value of business information online from the aspects of information, user and communication effect based on the current literature and marketing framework, and uses analytic hierarchy process to determine the weight of each level index, and establishes the fuzzy comprehensive evaluation model of the value of business information online. Finally the case study shows that it is reasonable and credible to evaluate the value of business information online with analytic hierarchy process and fuzzy comprehensive evaluation.

Keywords: e-commerce; business information online; analytic hierarchy process; fuzzy comprehensive evaluation.

1 Introduction

In the modern business activities, the importance of business information online has been paid more and more attention by small and medium-sized enterprises (SMEs). The business information online is not only the basis of making network marketing decisions and plans for enterprise, but also plays a very important role in the strategic management, market research and development of new products for enterprise. How to collect business information in the chaotic, decentralized, dynamic mass network information, and avoid the waste of unnecessary human, material and financial resources, and effectively evaluate the value of business information online, has become one of the essential capabilities for modern enterprise managers and operators.

During the exploitation and utilization of business information online, the value of business information online must be paid special attention. Thus evaluating the value of business information online has important significance which mainly reflects in the following two areas: (1) Get market opportunities and improve the competitiveness of enterprises. (2) Increase economic benefit and promote the construction of enterprise information.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 77–85, 2010.

In the network world, a large quantity of information will be released and published daily and hourly. The information, which posted on the enterprise websites, personal websites and the service platform of network information, has already constituted a complex system. This paper provides an evaluation index system of the value of business information online from the aspects of information, user and communication effect based on the current literature and marketing framework, then uses analytic hierarchy process to determine the weight of each level index, and establishes the fuzzy comprehensive evaluation model of business information online. Finally the case study shows that it is reasonable and credible to evaluate the value of business information online with analytic hierarchy process and fuzzy comprehensive evaluation.

2 Literature Review

Now, there are still few evaluation results of business information online. But few conceptual papers have offered systematic and comprehensive frameworks to assess and evaluate information value. Most research perspectives on the topic could be summarized as follows:

One of the perspectives is based on the authentication of information utility. Huanmin Xi, Ke Mou and Yongxing Peng in China suggested that the information value in decision-making reflected economic effects in the process of transforming a risky or uncertain decision-making into a deterministic decision process[1-2]. Yi Tang has certified information utility from the aspect of economic benefits[3].

The second perspective is based on the information utility and satisfaction of information users. Weisi Zhong, Jingxu Liu and Hongwei Li in China conducted the research on measurement from the perspectives of satisfaction and information utility based on the personalized information users[4].

The third perspective is based on the content and editing quality of information. Shuli Liu, Zhongeng Han, Liuyong Song and Shuqing Zhou in China evaluated information value from content and editing quality of information[5].

The fourth perspective is based on the attention of user and communication value. JinXing Hao in China put forward information value of lateral attention model from the cognitive and psychological characteristics of network information[6].

3 Construction of Evaluation Index System

3.1 Precondition

The value of business information online is dynamical, which would be changing with technology, users' behavior and external environments. Thus, this research is conducted based on the following preconditions.

- No great technology innovation on the Internet.
- No change of the users' behavior.
- A stable users' demand on business information.

3.2 Principle of Establishment of Evaluation Index System

As the value of business information online is dynamical, it will have certain difficulties to assess and evaluate the value. The establishment of evaluation index system of the value of business information online largely determines the validity of survey results and dependability. Thus, when we design the evaluation index system, the following principles should be followed: consistency of index with evaluation goal, compatibility of indexes in the same system, relative independence of each evaluation index, measurability, integrity and feasibility.

3.3 Establishment of Evaluation Index System

In order to reflect the actual situation of the value of business information online and the advantages and efficiency of business information online, and according to the characteristics of business information online, the value evaluation index system is constructed from the aspects of information, user and communication effect based on the current literature and marketing framework, as shown in Table 1.

Table 1. The evaluation index system for the value of business information online

	One-level indexes	Two-level indexes	Three-level indexes
The value of business information online (A)	User (B_1)	Demand (C_{11})	Practicality Importance Economy Interactivity Security
		Own factors (C_{12})	Operability
	Information (B_2)	Source (C_{21})	Authority
		Content (C_{22})	Authenticity Timeliness Accuracy Compliance Confidentiality Novelty Extensibility
		Editing quality (C_{23})	Normalization Simplicity Comprehensiveness Clarity
	Communication effect (B_3)	Acceptance effect (C_{31})	Click rate Response
		Satisfaction degree (C_{32})	Expectations Impressibility

User

The value of business information online refers to the users that business information online can meet the information needs of users. The information needs will determine users to judge the information value, and the value of business information online can be realized by the users. For the aspect of user characteristics, this paper constructs the indexes of demands and their own factors.

For the aspect of user demands, in the increasingly fierce market competition environment, the business information to enterprise are particularly urgent. Therefore, the value of business information online is also embodied in its ability to meet the information needs of information users. The value is mainly manifested in the following aspects It can provide personalized services, obtain important information at the lowest cost and make the information obtain maximize effectiveness, interact with the information users, and so on. The index of user demands can be measured by practicality, importance, economy, interactivity and security.

For the aspect of user own factors, the business information online is complicated and disorderly, and information users have individual differences, therefore, their ability to access to information, analyze information and process information is all very different. The information utility has direct relationship with the information users. The index of user own factors can be measured by operability.

Information

In order to judge the quality and information value, it is necessary to analyze the information source and judge the quality of information content. Thus, for the aspect of information characteristics, this paper constructs the indexes of source, content and editing quality.

For the aspect of information source, the source of business information online is the basic starting point of processing information, and is also an important reference of evaluating information value. The interactive, real-time and openness of network makes the source of business information diversify. Business information from different sources has different qualities. In general, the quality of information released by the authority of the institutions or well-known institutions is more reliable. The index of information source can be measured by authority.

For the aspect of information content, information content refers to the value of information content related to business activities, and it is the key attributes to information value. The index of information content can be measured by authenticity, timeliness, accuracy, compliance, confidentiality, novelty and extensibility.

For the aspect of editing quality of information, the value of business information online is also closely related to the editing quality. The information editor should accurately convey the results of collection, research, and its quality directly influences the information utility. The index of editing quality of information can be measured by normalization, simplicity, comprehensiveness and clarity.

Communication effect

Dissemination of business information online not only means the dissemination of information content, but also the dissemination of its commercial value. Thus, for the aspect of communication effect, this paper constructs the indexes of acceptance effect and user satisfaction.

For the aspect of acceptance effect, the value of business information online is also closely related to the acceptance effect. The acceptance effect of information is the target audience reaction to the dissemination of information. The index of acceptance effect can be measured by click rate and response.

For the aspect of user satisfaction degree, it refers to the satisfaction degree of information service for information users. It is the functions between the expectations of information utility before receiving information service and the perceived usefulness of information after receiving information. Thus, the index of user satisfaction degree can be measured by expectations and impressibility.

4 Comprehensive Evaluation

In the value evaluation of business information online, the analytic hierarchy process is used to determine the weight of each level index, and the fuzzy comprehensive evaluation is utilized to judge the value of business information online.

4.1 Establishment of Factor Set and Evaluation Set

Establishment of factor set *U*

Factor set is made up of elements that affect the judgment objects. According to the analysis above, there are seven single factors affecting the value of business information online, and they can be divided into two tiers. The factor set can be established as $U=\{U_1,U_2,U_3\}$, and the single-factor sets are $U_1=\{u_{11},u_{12}\}$, $U_2=\{u_{21},u_{22},u_{23}\}$, $U_3=\{u_{31},u_{32}\}$. U_1, U_2, U_3 respectively represents user, information and communication effect. U_1 is constructed by the indexes of user demands and user own factors, U_2 is constituted by the indexes of source, content and editing quality, U_3 is formed by the indexes of user acceptance effect and user satisfaction degree.

Establishment of evaluation set *V*

Evaluation set is made of all kinds of total judgment results given by judges as elements. The evaluation set V of the value of business information online can be established as $V=\{V_1,V_2,V_3,V_4,V_5\}$, in which five evaluation results are excellent, good, moderate, common and bad.

4.2 Establishment of Weight Set W

Every factor has different importance degree. To reflect the differences, every factor U_i is endowed with corresponding weight w_i. And the set $W=\{w_1,w_2...w_n\}$ which consists of weights is called factor weight set.

Establishment of the multi-level evaluation model

First, complex problems break down into several elements [7] and different elements are divided into several groups. Then we establish a multi-level evaluation model based on the group status.

Establishment of the comparison judgment matrix

Membership between the up-down hierarchy members is determined after we establish the multi-level evaluation model. We carry out the pairwise comparison between elements in each hierarchy of the multi-level model for the correlative up-level element, and then establish a series of judgment matrixes. According to Table1 above, we structure the comparison judgment matrix $A-B_i$ (i=1,2,3) as follows:

$$A-B_i=\begin{bmatrix}1 & 1/4 & 1/2\\ 4 & 1 & 3\\ 2 & 1/3 & 1\end{bmatrix}$$

Calculation of elements' relative weights in single criterion

This paper uses characteristic root method to compute collating weight vector. We suppose that the max characteristic root of judgment matrix is $\lambda_{\max}$, and the corresponding characteristic vector is *W*. The methods of *W* and $\lambda_{\max}$ are as follows:

$$w_i=\sqrt[n]{\prod_{j=1}^{n}b_{ij}}\Big/\sum_{i=1}^{n}\sqrt[n]{\prod_{j=1}^{n}b_{ij}}\quad (i=1,\cdots,n)\ . \tag{1}$$

$$\lambda_{\max}=\sum_{i=1}^{n}\frac{\sum_{i=1}^{n}b_{ij}w_i}{nw_i}\quad (i=1,\cdots,n)\ . \tag{2}$$

Based on Formulation (1) and (2), we calculate and get the $\lambda_{\max}$ and W_B of comparison judgment matrix $A-B_i$: $\lambda_{\max}$ =3.0183, W_B =(0.1365,0.6250,0.2385). *W* is the weight set of *B*-level elements for the general goal.

Consistency check

To make sure that the decision-making process is scientific, consistency check of $\lambda_{\max}$ is necessary. Checking process is as following:

$$CI=(\lambda_{\max}-n)/(n-1)\ . \tag{3}$$

$$CR=CI/RI\ . \tag{4}$$

In the Formulation (3) and (4), *CI* stands for coincidence index, *CR* represents coincidence rate and *RI* is random coincidence index.

When *CR*<0.1, we consider that judgment matrix has a good consistency, or else we should adjust the values of elements in judgment matrix.

Based on Formulation (3), we calculate and get *CI of* comparison judgment matrix $A-B_i$: *CI*=0.0091, when *n*=3 and *RI*=0.58.

Based on Formulation (4), we calculate and get *CR of the* comparison judgment matrix $A-B_i$: *CR*=0.0158<0.10. This indicates that the judgment matrix has a satisfying consistency.

Similarly, we can establish the judgment matrixes of *C*-level elements for correlative *B*-level elements, and also calculate all weights of evaluation indexes of business information online, as shown in Table 2.

Table 2. Relative weights of the value of business information online evaluation indexes

One-level indexes	Weight	Two-level indexes	Weight
User	0.1365	Demand	0.8000
		Own factors	0.2000
Information	0.6250	Source	0.1168
		Content	0.6833
		Editing quality	0.1998
Communication effect	0.2385	Acceptance effect	0.3333
		Satisfaction degree	0.6667

Calculation of combination weight of each level element

To get the weights of all elements of each level for the overall objective, it is necessary to judge the value of CR. If $CR \geq 0.1$, we should assemble the calculation results of the third step properly and check the total judgment consistency. We do this step basipetally. The final results indicate the relative weight of decision-making priority sequence and the judgment consistency check of the whole hierarchical model.

According to the results of Table2, we can calculate and get combination weight of seven evaluation indexes: *W*=(0.1092,0.0273,0.0730,0.4271,0.1249,0.0795,0.1590).

4.3 Fuzzy Evaluation

First, experts evaluate from the single element of factor set *U* and determine the degree of membership that the evaluation objects rely on the elements of factor set. Then, we establish the total evaluation matrix consisting of evaluation sets of n elements. It is usually expressed as *R*[8].

After we get values of *W* and *R*, we can do fuzzy mapping to have a comprehensive judgment. The mathematical model of fuzzy comprehensive evaluation is shown as:

$$B = W \cdot R = \left(v_{b1}, v_{b2}, \ldots, v_{b5}\right) . \tag{5}$$

In the Formulation (5), v_{bi} $(i = 1,2,\ldots,5)$ respectively represents the membership of different evaluation results.

5 Application Example

Evaluating factor set *U* is made up of seven factors influencing the value of business information online. Weight set *W* is established with seven evaluation indexes: *W*=(0.1092,0.0273,0.0730,0.4271,0.1249,0.0795,0.1590). Evaluation set *V* is established with five evaluation results for the factors: excellent, good, moderate, common and bad.

Take an enterprise as example, this enterprise collected three pieces of information from the network. According to experts' test data of the value of business information online, we establish estimation matrix. The estimation matrix of information 1 is as shown in Table 3.

Table 3. Estimation matrix R of information 1

	Excellent	Good	Moderate	Common	Bad
Demand	0.10	0.30	0.50	0.10	0
Own factors	0.10	0.30	0.40	0.20	0
Source	0.10	0.40	0.50	0	0
Content	0.10	0.50	0.40	0	0
Editing quality	0.10	0.40	0.40	0.10	0
Acceptance effect	0.20	0.20	0.50	0.10	0
Satisfaction degree	0	0.10	0.60	0.30	0

Based on Formulation (5), we can calculate and get the comprehensive judgment of information 1: B=(0.0921,0.3655,0.4580,0.0845,0).

Similarly, we establish the estimation matrix of information 2 and information 3 and calculate and get their comprehensive judgment, as shown in Table 4.

Table 4. Estimation result of three pieces of information

	v_{b1}	v_{b2}	v_{b3}	v_{b4}	v_{b5}
Information 1	0.0921	0.3655	0.4580	0.0845	0
Information 2	0.1250	0.4191	0.36749	0.0884	0
Information 3	0	0.1103	0.30341	0.5041	0.0719

According to the principle of maximum degree of membership, the value of information 2 is the second level, namely, that is good; the value of information 1 is moderate; but the value of information 3 is common.

6 Conclusions

This paper adopts the method of analytic hierarchy process and fuzzy evaluation to establish the fuzzy comprehensive evaluation model of the value of business information online, in order to avoid the effect of individual subjective judgment and favoritism on the result of business information online evaluation.

According to the results, fuzzy comprehensive evaluation is a reasonable and feasible method to evaluate the value of business information online, and it can be used widely in the value evaluation of business information online.

In summary, the value evaluation of business information online has attracted many people's attention. We hope that the value evaluation index system of business information online and preliminary evaluation results will help to further research in this area.

Acknowledgments. This paper is supported by the Project of SMEs Development Research Center Platform Construction (PXM2010_014209_098878).

References

1. Xi, H., Mou, K.: Utility Theory and Information value of the Hotel Business Decision-making. Journal Beijing Second Foreign Language Institute 1, 53–59 (1994) (in Chinese)
2. Peng, Y.: Measure the value of risk decision-making information. Statistics and Decision 9, 32–33 (1995) (in Chinese)
3. Tang, Y.: On the Economic Efficiency of Information and Evaluating Method (in Chinese). The Library Journal of Henan 23, 18–20 (2003)
4. Zhong, W., Liu, J., Li, H.: The Study of the Information Value Measurement Based on Information Effectiveness and Information User Satisfaction Degree (in Chinese). Journal of Information 1, 92–93 (2007)
5. Liu, S., Han, Z., Song, L., Zhou, S.: Comprehensive Evaluating Model for the Value of Information. Journal of Information Engineering University 8, 118–120 (2007) (in Chinese)
6. Hao, J.: Attention Model for Measuring Information Value. Journal of the China Society for Scientific and Technical Information 5, 618–625 (2003) (in Chinese)
7. Saaty, T.L.: The Analysis Hierarchy Process. McGraw Hill, New York (1980)
8. Buckley, J.J.: Ranking Alternatives Using Fuzzy Numbers. Fuzzy Sets and Systems 15, 21–31 (1985)

A Clustering Algorithm Based on Matrix over High Dimensional Data Stream

Guibin Hou[1], Ruixia Yao[1], Jiadong Ren[1,2], and Changzhen Hu[2]

[1] College of Information Science and Engineering Yanshan University
066004 Qinhuangdao, China
[2] School of Computer Science and Technology Beijing Institute of Technology
100081 Beijing, China
houguibin@portqhd.com, yrxstar@hotmail.com,
jdren@ysu.edu.com, cz_hu@sina.com

Abstract. Clustering high-dimensional data stream is a difficult and important issue. In this paper, we propose MStream, a new clustering algorithm based on matrix over high dimensional data stream. MStream algorithm incorporates a synopsis structure, called GC (Grid Cell Structure), and grid matrix technique. The algorithm adopts the two-phased framework. In the online component, the GC is employed to monitor one-dimensional statistics data distribution of each dimension independently. Sparse GCs which need to be deleted are checked by predefined threshold. In the offline component, it is possible to tracing multi-dimensional clusters by dense GCs which are maintained in the online component. Grid matrix technique is introduced to generate the final multi-dimensional clusters in the whole data space. Experimental results show that our algorithm has the flexible scalability and higher clustering quality.

Keywords: high-dimensional; clustering; grid matrix.

1 Introduction

In recent years, with explosion of streaming data, data mining in data streams has become a significant issue [1]. Cluster plays an important role in data mining field as an important traditional data mining method. It can be used to address many problems, such as intrusion detection [2], mining web documents [3].

Recently, Aggarwal devised a clustering framework for stream called CluStream [4]. Data records are compressed in micro-clusters which are stored at snapshots in time which follow a pyramidal pattern. Based on the two-phased framework in CluStream, Chen presented D-Stream [5] which is grid-based clustering algorithm. In the online component, each input data record is assigned into a corresponding grid. In the offline component, the grid density is computed and the final clusters are obtained. Cao presented DenStream [6] based on density can find arbitrarily shape clusters over damped window of data stream with noises.

However, a lot of stream data is high-dimensional in nature. The above algorithms are unable to deal with clustering over high dimensional data stream. Considering the sparsity of the data under the high-dimensional circumstance, all pairs of points tend

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 86–94, 2010.

to be almost equidistant from one another based on the distance measure. The distance measure over full dimensions can not be applied to the clustering of high-dimensional data streams.

In order to cluster on high-dimensional data streams, much recent work has adopted the methods of projected clustering technique and grid-cell clustering technique. Aggarwal presented a new high-dimensional projected data stream clustering method, called HPSream [7]. The technique of projection based clustering methodology is introduced to clustering high-dimensional data streams. In order to embody the importance of the latest data HPStream uses fading cluster structure to store the information of historical data. The algorithm is effective in finding clustering over high-dimensional data streams, but it is difficult to predetermine the value of l (the average number of projected dimensions) in real application. WSCStream [8] is an algorithm clustering on subspace to reduce the number of dimension associated with clusters. A dimensional weight matrix is proposed to calculate dynamic subspace of clusters. Recently, Aggarwal studied the problem of projected clustering of uncertain data streams in [9]. The results show that the projected technique can be effectively adopted in the context of uncertain data streams. Lu presented a grid-based subspace clustering algorithm, GCHDS [10], for clustering high-dimensional data streams. The arriving data are summarized into a grid data structure. Useful dimensions are selected to construct a subspace in which the clustering process is performed. Park [11] devised an algorithm equipped the capability to deal with high dimensional data stream. Sliding list was applied to maintain the statistic information of each dimension. Based on the sibling list, a cluster-statistics tree was proposed to monitor the density of multi-dimensional clusters. But it is difficult to predefine the sequence of dimensions to tracing the multi-dimensional cluster.

In order to clustering of high dimensional data stream, CluStream clustering framework is adopted. In the online component, GC structure is proposed to monitor the data distribution information for each dimension. In the offline component, grid matrix which is comprised of dense GCs is able to traverse from arbitrary position (row and column) and arbitrary multi-dimensional clusters can be found in the whole data space.

The rest of the paper is organized as follows: Section 2 discusses the basic definitions and concepts in our algorithm. In Section 3, the algorithm based on grid matrix will be introduced. Our performance study on real data sets is reported in Section 4. Finally, Section 5 concludes the paper.

2 Basic Concepts and Definitions

We assume that $S=S_1 \times S_2 \times \ldots \times S_d$ represent the data space of a data stream. The set of data points E= (e_1, e_2, … , e_n) is arriving at time stamps $t_1 \ldots t_k \ldots$, wherein each ei is a d-dimensional data record, denoted as e_i= (e_{i1}, e_{i2}, …, e_{id}).

Clustering patterns of data stream usually change with time. The latest data is more significant than old in a real stream. To achieve this, the density of an old data is decreased. If a data point e arrives at time t_c, the density fading function at time t is $D(e,t) = \lambda^{t-t_c}, 0 < \lambda < 1$. In particularly, when $t = t_c$ the density is $D(e,t) = 1$.The fading function is widely used in temporal applications where it is desirable to gradually discount the history of past behavior.

To provide flexible scalability on the number of dimensions as well as the size of data sets, grid cell is explored to maintain synopsis data distribution of each dimension. We partition the d-dimensional spaces S into irregular grid cells. Suppose for each dimension, its space S$_i$, $i = 1, 2, \cdots d$, is divided into partitions as $S_i = S_{i,1} \bigcup S_{i,2} \bigcup \cdots \bigcup S_{i,p}$, p is a variable number according the data distribution on each dimension. A grid cell for the space S$_i$, m (1≤m≤p) is defined as follows.

Definition 2.1 (Gird Cell) For a group of close points $e_{1j}, e_{2j} \cdots e_{nj}$ with time stamps $t_1 \ldots t_k \ldots$, a term GC (CF1, CF2, D, I, T) is used to denote the distribution statistics of a grid cell which is defined by an interval $I = [s, f)$ in the jth dimension. The definition of each of these entries is as follows:

- For each dimension, the weighted sum of the data values is maintained in $CF1 = \sum_{j=1}^{n} \lambda^{t-t_{lj}} e_{ij}$. The t$_{lj}$ is the creating time of this GC structure in jth dimension.
- For each dimension, the weighted squared sum of the data values is maintained in $CF2 = \sum_{j=1}^{n} \lambda^{t-t_{lj}} e_{ij}^2$.
- For all data points in interval I, the sum of decayed density is maintained in $D = \sum_{j=1}^{n} \lambda^{t-t_l}$. We define that if density of grid cell is above β (a predefined threshold), this grid cell is dense, or this grid is sparse.
- The equation I=[s, f) is the interval of grid cell. We also define the center of the grid cell is $c = CF1 / \mathrm{D}$. The radius of grid cell is defined as $r = \sqrt{CF2 / \mathrm{D} - (CF1 / \mathrm{D})^2}$, $I.s = c - r, I.f = c + r$. There exists ordering relationship among grid cells according the interval I. e.g. if GCi.I<GCj.I, then GCi(CF1,CF2,D,I,T)<GCj(CF1,CF2,D,I,T) .
- T is the latest update time of this grid cell.

The data distribution of the jth dimension in data space S is described by a set of non-intersecting grid cells {GC1, GC2, … ,GCn} where $GC1.I \bigcap GC2.I = \varnothing$. A set of GC maintain the distribution statistics of those data points in the jth dimension.

2.1 Grid Cell Maintenance

In order to discover clusters in an evolving data stream, we maintain a group of GC for each dimension in online. This is based on the observation that most of the new points belong to existing clusters, therefore can be absorbed by existing GC. When a new data point e arrives, the procedure is described below.

(1) If there is a grid cell GC of which GC.*I* is corresponding the value of e$_{ij}$, then e$_{ij}$ is mapped to the grid cell GC, and each of entries are updated.

(2) If there isn't a grid cell GC of which GC.*I* is corresponding the value of e$_{ij,}$ then a new GC is created to maintain the data point e$_{ij.}$

GC can be maintained incrementally. Suppose a new data point is mapped to the grid cell at time t, and suppose the time when GC receives the last data points is $t_{c,}$ then the grid cell is updated as follows:

$$D = \lambda^{t-t_c} D(g, t_c) + 1 \ . \tag{1}$$

$$CF1 = \lambda^{t-t_c} CF1 + e_{ij} \ . \tag{2}$$

$$CF2 = \lambda^{t-t_c} CF2 + e_{ij}^2 \ . \tag{3}$$

The interval I is also updated by calculating the new radius and center for the new data point follows CF1 and CF2. These GCs are organized follow the sequence of the entries of the interval I. If there is not an intersection in j^{th} dimension between grid interval $[I_{ij}.\ s, I_{ij}.\ f)$ and another present grid, a new grid cell GC $[I_{ij}.\ s, I_{ij}.\ f)$ is added, and the new GC is insert into the corresponding position. GCs also may have overlapping. There are three cases for adjusting the GCs as follows. The distribution statistics of the j^{th} adjusted GCs are initialized by the normal distribution function, the detail has been discussed in [11].

(1) If $I_{i\,j}.\ s< I_{i-1,\,j}.\ f$, new grid cell $[I_{i-1,\,j}.\ f,\ I_{i\,j}.\ f)$ is added.
(2) If $I_{i\,j}.\ f>I_{i+1,\,j}.\ s$, new grid cell $[I_{i\,j}.\ f,\ I_{i+1,\,j}.\ s)$ is added.
(3) If $I_{i\,j}.\ s< I_{i-1,\,j}.\ f$ and $I_{i\,j}.\ f>I_{i+1,\,j}.\ s$, a new grid cell $[I_{i-1,\,j}.\ f,\ I_{i+1,\,j}.\ s)$ is added.

2.2 Grid Detection

The density of any GC is constantly changing. A dense grid may degenerate to a sparse grid if it receives a little new data for a long time. We have found that it is unnecessary to update the grids at every time step. After a period of time, the density of each grid should be detected. On the other hand, most new created GC its density is relatively small compared with existing GC at the initial stage. It is necessary to provide some time for a new GC to grow into a dense GC. Therefore, we define the time interval T_o and sparse density threshold η which are similar to these in [6]. The main differences between them are the definition of dense GC.

Suppose the time interval is T_o to check the density of the dense grid cells. According to the equation $\lambda^{T_o}\beta + 1 = \beta$, we yields the minimum time interval needed for a dense grid to degenerate to a sparse grid is

$$T_o = \left\lceil \log_\lambda \frac{\beta - 1}{\beta} \right\rceil \ . \tag{4}$$

For a GC, when its density becomes less than or equal to η, its possibility of becoming a dense grid cell in the near future is considered to be low enough to disregard it. The sparse density threshold is defined as:

$$\eta = \frac{\lambda^{t+T_o-t_c} - 1}{\lambda^{T_o} - 1} \ . \tag{5}$$

2.3 Grid Matrix

Each dense GC is considered as a node in a d-dimensional data space. In order to manage the dynamically varied configuration of GC in the entire data space of the dimension efficiently, the GCs are structured by a grid matrix GM defined as follows.

Definition 2.2 (Grid Matrix) The vertex in the i-th row and j-th column is corresponding to a vector, denoted as <D(j), next-pre>. D(j) is the j-th density of grid cell in the i-th dimension. The next-pre is a pointer which directs to the dense GC in the next dimension. If the dense GCi in j-th dimension has not union with GCj in the next dimension, next-pre is set as null. The number of row is equal to the maximum number of dense GC among the d dimension. The number of column is equal to the number of dimension.

The density of grid cell g is obtained by the equation of $D(g) = GC_1 \wedge GC_2 \cdots \wedge GC_d$, which is the minimum of the density of all GC comprised the gird cell. If the density of grid cell meets *D(g)>β,* then the grid is dense. Fig 1 is the grid structure in two dimensional data space. These gird cells GC(1,2) GC(2,2),GC(3,3), GC(5,1), GC(6,3) are dense.

Figure 2 is a grid matrix in a two-dimensional data space. The final cluster is obtained through traversing the matrix GM. In the GM, it is benefit to traverse the matrix from the any row and column. For example, we scan GM from the position of 1-th row and 1-th column. There are five 2-dimensional clusters in the Fig 2, such as D(1)->D(2), D(2)->D(2),D(3)->D(3),D(5)->D(2),D(6)->D(3). It is possible for user to obtain arbitrary k-dimensional cluster which k is changed from 1 to d.

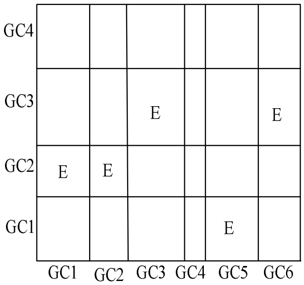


Fig. 1. Grid structure of two-dimension

$$\begin{array}{c|ccccc} & 1 & 2 & 3 & 4 & 5 \\ \hline 1 & D(1),2 & D(2),2 & D(3),3 & D(5),2 & D(6),3 \\ 2 & D(1),0 & D(2),0 & D(3),0 & 0 & 0 \\ 3 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & & & \\ d & & & & & \end{array}$$

Fig. 2. Grid Matrix

3 Description of MStream Algorithm

Our algorithm adopts the two-phased clustering framework as well as CluStream algorithm [4]. In the online component, we maintain GC structure for each dimensions, and update GCs with the data stream arriving. In each time periods T_o, it is necessary to check these GCs for updating the density and deleting the sparse GCs. In the offline component, grid matrix is constructed to denote the relationship between two adjacent dimensions. When the clustering request arrives, traversing the grid matrix follows the next-pre point to get the final clustering results.

For all the dense GCs whose density meet D(g)>β, they are denoted as the input in algorithm 1. If the grid formed by the GCi which is in the d-th dimension and GCj in the d+1-th dimension is dense, then the next-pre is equalize to j. Algorithm 1 shows how to construct the Grid Matrix GM.

```
Algorithm 1 GetGridMatrix (D(g),d)
Input: D(g):the density of dense GCs, d: the number of dimensions;
Output: Grid Matrix: GM;
for (i=1;i<=d;i++) {
   for (j=0;j<GetNumber(GCs in the i$^{th}$ dimension); j++)
{
   for(k=0; k<GetNumber(GCs in the i+1$^{th}$ dimenson);k++) {
      if (D(j)∧D(k)>  )
           next->pre=k;
      else  next->pre=0;}
Return GM;}}
```

When a new data point e arrives with the timestamp t, we first check the existing GCs for each dimension of e. If there is a GC of which the interval I corresponding to e_{ij}, then e_{ij} is mapped to this GC, and GC is updated; else, a new GC is created in this dimension if there isn't a GC corresponding to e_{ij}. All existing GCs are adjusted to follow the sequence defined in definition 2.1. The GC is deleted if its density is less than the sparse density threshold after the period of T_o. The corresponding record in Grid Matrix is also deleted. Subsequently, the Grid Matrix GM is constructed for all the dense GCs. When the clustering request arrived, the final multi-dimensional clusters are obtained according to row and column vectors in Grid Matrix. The description of MStream algorithm is as follows:

```
Algorithm MStream
Input: d-dimensional data stream, λ, β;
Output: grid cell g marked with clusterID.
$T_o = \left\lceil \log_\lambda \frac{\beta-1}{\beta} \right\rceil$
Initialize Grid Matrix GM={0};
for each data point e in data stream{
   for each dimension of data point e{
      if e∈ a existing GC then
          Update entries in GC;
      else { Create a new GC in d-th dimension;
             Adjust the GCs in d-th dimension;}
      if (t mod $T_o$=0) then {
              $\eta = \frac{\lambda^{t+T_o-t_c}-1}{\lambda^{T_o}-1}$;
      if GC.D<η  then
          Delete GC from the j$^{th}$ dimension;}}}
GetGridMatrix (GC.D,d);
if (clustering request arrived)
    Traverse GM follows the next-pre pointer;
Output the dense GC marked clusterID;
```

4 Experiments

In order to analyze the performance of proposed method, we compare the algorithm with HPStream and CS-Tree algorithms, both of them are used to deal with high-dimensional data streams. All experiments are conducted on a 2.4Ghz Pentium PC with 512GB memory space running on the Windows XP platform. The real testing dataset is the KDD CUP-99 network intrusion detection stream data. It contains a total of 494,020 data records and all the 34 continuous attributes are used for clustering. In addition, we also use synthetic data sets to test our algorithm. The synthetic data sets are generated by the data generator used in [8]. The domain of each dimensional value is ranged over [0,100) and the value of each data elements is randomly varied. The number of dimension is varied from 50 to 90. Data points are looked up one by one in sequence to simulate the environment of data stream.

The input parameter of the MStream algorithm are set as λ=0.998, β = 5. The parameter β can be set by the average density of points in each grid cell. Generally, we should choose a smallest acceptable. Instead of using SSQ, The clustering accuracy is evaluated by the cluster purity which is defined as the average percentage of the dominant class label in each cluster.

According to the Fig 3, we conclude that our algorithm has higher clustering quality than HPStream and CS-Tree algorithm. The cluster purity of MStream is always more than 90%. This is because MStream use GC to accurately capture the distribution characteristic of the data points on each dimension. However, it is difficult for users to predefine the average number of projected dimensions *l* in HPStream. Fig 4 also shows the clustering quality of MStream by varying the number of dimensions on the synthetic data set. When the number of dimensions increases, we only need to increase the number of GCs and sibling entries respectively. So MStream and CS-Tree have higher clustering quality than HPStream.

We test the scalability against dimensionality. The number of dimension is set as 50, 60, 70, 80 and 90, and we set the stream speed as 200 points per time unit. The Fig 5 shows that MStream has linear increase in runtime for data sets, and also has the more smooth growth than HPStream and CS-Tree algorithms. This is because when

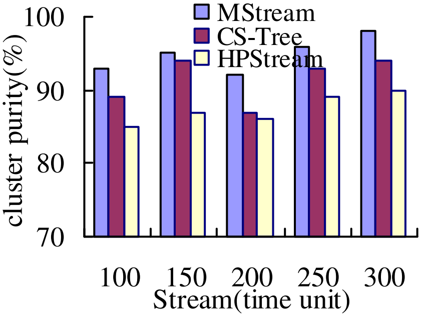


Fig. 3. Quality comparisons for the Network Intrusion dataset

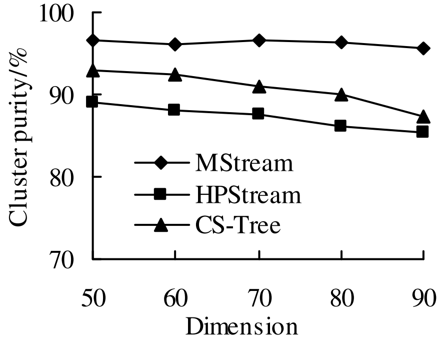


Fig. 4. Quality comparisons by varying dimensionality

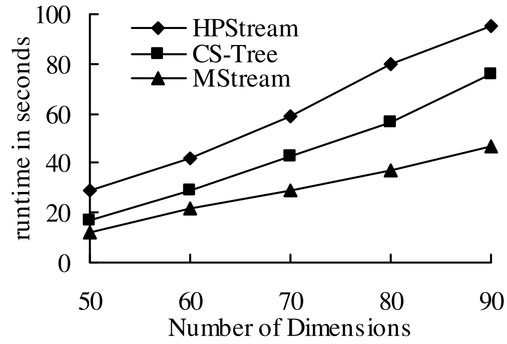


Fig. 5. Scalability with dimensionality

the number of dimensions increases, MStream only needs to increase the number of GCs and the size of Grid Matrix, while CS-Tree needs to increase the number of sibling lists and reconstruct CS-Tree. The maintenance operation of GCs and Grid Matrix costs less time, compared with construction of trees. In HPStream, computing the projected dimensions will cost a lot of time when the dimensionality increases. So, the scalability of MStream with dimensionality is higher than those of HPStream and CS-Tree.

5 Conclusion

We propose MStream, a new clustering algorithm based matrix over high dimensional data stream. Instead of clustering in projected subspace, MStream deals with the dimensionality one by one. MStream adopts two-phased clustering framework. In the online component, GC structure is introduced to maintain the synopsis information of each dimension. In the offline component, multiple-dimensional clusters are obtained by way of traversing grid matrix. The final cluster result is a string linked by the pointer of next-pre in GM. This method has flexible scalability on the number of dimensions as well as the size of a data set. Experiments demonstrate that our algorithm is able to cluster the high dimensional data streams and gets the higher clustering quality.

Acknowledgments. This work is supported by the National High Technology Research and Development Program ("863"Program) of China under Grant No. 2009AA01Z433, and the Natural Science Foundation of Hebei Province P. R. China under Grant No. F2010001298. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Babcock, B., Babu, S., Datar., M., Motwani, R., Widom, J.: Models and issues in data stream systems. In: The twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Madison, Wisconsin, USA, pp. 1–16 (2002)
2. Park, N., Oh, S.H., Lee, W.: Anomaly intrusion detection by clustering transactional audit streams in a host computer. J. Information Sciences 180, 2375–2378 (2010)
3. Sambasivam, S., Theodosopoulos, N.: Advanced data clustering methods of mining Web documents. J. Issues in Informing Science and Information Technology 3, 563–579 (2006)
4. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: 29th International Conference on Very Large Data Bases, Berlin, Germany, pp. 81–92 (2003)
5. Chen, Y., Tu, L.: Density-Based Clustering for Real-Time Stream Data. In: 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, pp. 133–142 (2007)
6. Cao, F., Ester, M., Qian, W.N., Zhou, A.: Density-based clustering over an evolving data stream with noise. In: SIAM Conference on Data Mining, pp. 326–337 (2006)
7. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for projected clustering of high dimensional data stream. In: The 13th International Conference on Very Large Data Bases, Toronto, pp. 852–863 (2004)
8. Ren, J., Li, L., Hu, C.: A Weighted Subspace Clustering Algorithm in High- Dimensional Data Streams. In: Fourth International Conference on Innovative Computing Information and Control, Taiwan, pp. 631–634 (2009)
9. Aggarwal, C.C.: On high dimensional projected clustering of uncertain data streams. In: IEEE International Conference on Data Engineering, pp. 1152–1154 (2009)
10. Lu, Y.S., Sun, Y.F., Xu, G.P., Liu, G.: A grid-based clustering algorithm for high-dimensional data streams, pp. 824–831. Springer, Heidelberg (2005)
11. Lee, J.W., Park, N.H., Jee, W.S.: Efficiently tracing clusters over high-dimensional on-line data streams. J. Data & Knowledge Engineering 68, 362–379 (2009)

A Dynamic Dispatcher-Based Scheduling Algorithm on Load Balancing for Web Server Cluster

Liyong Bao, Dongfeng Zhao, and Yifan Zhao

School of Information Science and Engineering, Yunnan University
Kunming, 650091, P.R. China
lybao@ynu.edu.cn, zhaodf123@263.net,
shipzhaoyifan@yahoo.com.cn

Abstract. Effective Request distribution and load balance are the key technological means to guarantee higher-quality service in web server clusters system. Based on the ideas of dynamic load adaptation and priority service, this paper presents a new scheduling algorithm of load balance of web cluster servers. The theoretical model of this scheme is established with Markov chain and probability generating function. Mathematical analysis is made on the mean queue length and the mean inquiry cyclic time of the common queue and the key station. The findings of theoretical analysis and the experiments show that this algorithm optimizes the services of the system in time of load change input via adjustment of the times of the access gated, and strengthens the flexibility and fairness of multimedia transmissions in web cluster system.

Keywords: Web server cluster; load balance; dispatcher-based scheme; mean queue length; mean cyclic time.

1 Introduction

The popularity of the internet has increased dramatically in the past few years. World Wide Web is becoming one of the most effective ways in sharing information and obtaining services. However, with the explosive growth in Web traffic in business and other domains, the quality of service is declining. The overloading of Web servers and the delayed service because of network congestion have become the predominant problems in Web service[1-3]. Many researches have been done on improving Web performance and scalability. As is well known, a single high-function server can hardly meet the increasing demand for service[4]. Considering the situation, an effective solution to the problem lies with the application of Web server cluster technology so as to meet the demands of higher-quality service.

Cluster server must provide customers with a single system image (SSI). In other words, a customer's request is transparently redirected to clusters within the service nodes, allowing the user to take the cluster as a high performance server. Dispatcher-based approach is an important scheme technique. The dispatcher receives all requests from clients and distributes the requests to servers. Request routing among servers is transparent. There are several variations of dispatcher-based architectures using routing mechanisms: packet rewriting, packet forwarding, and HTTP redirection [5–7].

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 95–102, 2010.

Dispatcher-based architectures typically use simple algorithms to decide which Web server will manage incoming requests from a given client, as simple algorithms help minimize request processing. Examples of such simple algorithms include random selection, round robin, and least-connection. Dispatcher-based approaches are hampered by a single-decision entity, which can be a bottleneck with increased requests. Furthermore, in centralized scheduling, a failed dispatcher can disable the system (i.e., there is a single point of failure). Nevertheless the dispatcher can achieve fine-grained load balancing [8].

The key of Dispatcher-based scheme mechanism is to carry out the task allocation and scheduling. Task allocation depends on each server of the load information acquisition and processing techniques. How to obtain and collect the web servers load information to the benefit of task scheduling is the focus of research in load balancing system.This paper presents a new scheme of load balance of Web cluster servers. Based on queuing theory [9-11], this research establishes the dynamic polling scheduling system.

2 The Theoretical Model of the Scheduling Algorithm

The system is made up of one dispatcher, one key station h at the first level, and N common stations at the second level. The higher-priority key station applies an exhaustive-service policy with less time delay, which fits for the highly real-time performances. Meanwhile, the lower-priority common station uses the gated policy with one or two-times services, which is based on the dynamically adjustable input load, to facilitate non-real-time transmissions. When the network load increases, service quality can be maintained with the increase of service times, which also fulfills the requirements of the non-real-time performances for fairness and time delay. The first and the second stations receive alternatively the polling connection service from dispatcher.

2.1 The Mechanism and Variable Definition of the System Model

Polling services at each station include the following processes:

1) The dispatcher polls at the common station of i $(i=1,2,\cdots,N)$ at the time of t_n. The lower priority business information packets number of queuing waiting for transmission in the buffer at the i station is $\xi_i(n)$, when the high priority request information packet number of queuing waiting for transmission in the memory at the key station is $\xi_h(n)$. The dispatcher provides transmission service for the i station where the information packet in its queue is completed at the time of t_{n1}. Then, if there are new information packets coming into the memory of the i terminal station within the service time, the lower priority request information packet number of queuing waiting transmission in the memory at the i station is $\xi_i(n1)$, and the higher priority business information packet number of queuing waiting for

transmission in the buffer at the key station is $\xi_h(n1)$. According to the service regulations, the server continues to work for the i terminal station, completing information packets within its queue via gated service for the second time in the i station at the time of t_{n2}. The overall time of the dispatcher's service of the i lower priority queue is $v_i(n)$, with the information packets number of $\eta_j(v_i)$ when entering the j station $(j=1,2,\cdots N,h)$ at the time of $\eta_j(v_i)$.

2) After the three-times gated service and via the transformational time of $u_{i1}(n)$,the inquiry starts for the information packets queue waiting to be transmitted at the key station h, with the information packets number of $\mu_j(u_{i1})$ when entering the j station $(j=1,2,\cdots N,h)$ at the time of u_{i1}.

3) The key station h receives service at the time of $t_{\overline{n}}$, when the higher priority information packets number of queuing waiting for transmission in the buffer at the key station is $\xi_{ih}(\overline{n})$. After the completion of information grouping transmission within the key station via exhaustive service policy, the time of the server's service of the higher priority queue at the key station h is $v_h(n)$, with the information packets number of $\eta_j(v_h)$ when entering the j station $(j=1,2,\cdots N,h)$ at the time of v_h.

4) After the service of the key queue, there is a transformational time of $u_{i2}(n)$. Transmission service comes to the $i+1$ common station at the time of t_{n+1}, with the information packets number of $\mu_j(u_{i2})$ when entering the j station $(j=1,2,\cdots N,h)$ at the time of u_{i2}. Likewise, the lower priority business information packets number of queuing waiting for transmission in the buffer at the i $(i=1,2,\cdots,N)$ station is $\xi_i\ (n+1)$ at the time of t_{n+1}.

2.2 The Operational Conditions for the System

According to the operational process of the system, the operational conditions are defined as follows:

1) The information packets entering each common station follow the independent and identical distribution of probability, with the distributional probability generating function and mean value respectively as $A(z)$ and $A'(1)=\lambda$, and with the probability generating function and mean value of the information packets arriving at the key station h respectively as $A_h(z)$ and $A_h'(1)=\lambda_h$;

2) The time used by an information packet in the buffer of the common station when the dispatcher transmits services follows an independent and identical distribution of probability, with the distributional probability generating function and mean value respectively as $B(z)$ and $B'(1)=\beta$, and with the distributional

probability generating function and mean value of the transmission service time at the key station h respectively as $B_h(z)$ and $\beta_h = B'_h(1)$;

3) The random variable of the inquiry transformation time when polling from the i common station to the key station h follows an independent and identical distribution of probability, with the distributional probability generating function and mean value respectively as $R_{i1}(z)$ and $R'_{i1}(1) = \gamma_{i1}$. The random variable of the inquiry transformation time when polling from the $i+1$ common station to the key station h follows an independent and identical distribution of probability, with the distributional probability generating function and mean value respectively as $R_{i2}(z)$ and $R'_{i2}(1) = \gamma_{i2}$;

4) The random variable of the time for the exhaustive service of information grouping arriving at the key station at any time slot follows an independent and identical distribution of probability, with the distributional probability generating function as $F_c(z)$;

5) The system operates on the FCFS rule. There is enough capacity of the memory at each terminal station so that information grouping will not get lost.

2.3 Probability Generating Function of the System Status Variable

The system remains stable in the condition of $\sum_{i=1}^{N} \rho_i + \rho_h < 1$ ($\rho = \lambda\beta$, $i = 1,2,\cdots,\ N$). The system status can be described with the Markov chain which is non-periodic and ergodic. According to the operating process of the system, the probability generating function of the system status variable is shown as follows:

at the time of t_{n2}

$$G_{i2}(z_1, z_2, \cdots, z_N, z_h) = \lim_{t\to\infty} E\left[\prod_{i=1}^{N} z_i^{\xi_{n2}} z_h^{\xi_{n2}}\right]$$

$$= G_{i1}\left(z_1, z_2, \cdots, B_i\left(A_h(z_h)\prod_{j=1}^{N} A_j(z_j)\right), z_{i+1}, \cdots, z_N, z_h\right) \tag{1}$$

at the time of $t_{\bar{\mathrm{n}}}$

$$G_{ih}(z_1, z_2, \cdots, z_N, z_h) = \lim_{t\to\infty} E\left[\prod_{i=1}^{N} z_i^{\xi_i(\bar{\mathrm{n}})} z_h^{\xi_h(\bar{\mathrm{n}})}\right]$$

$$= R_{i1}\left(A_h(z_h)\prod_{j=1}^{N} A_j(z_j)\right) G_{i2}\left(z_1, z_2, \cdots, B_i\left(A_h(z_h)\prod_{j=1}^{N} A_j(z_j)\right), z_{i+1}, \cdots, z_N, z_h\right) \tag{2}$$

at the time of t_{n+1}

$$G_{(i+1)1}(z_1, z_2, \cdots, z_N, z_h) = \lim_{t\to\infty} E\left[\prod_{i=1}^{N} z_i^{\xi_i(n+1)} z_h^{\xi_h(n+1)}\right]$$

$$=R_{i2}\left(A_h(z_h)\prod_{j=1}^{N}A_j(z_j)\right)G_{ih}\left(z_1,z_2,\cdots,z_N,B_h\left(\prod_{j=1}^{N}A_j(z_j)F_c\left(\prod_{j=1}^{N}A_j(z_j)\right)\right)\right) \quad (3)$$

2.4 The Analysis of Mean Queue Length

Given that $g_i(j)$ is the average number of information packets number at the j station buffer when the i station starts service at the time of t_n, then

$$g_i(j)=\lim_{z_1,\cdots,z_N,z_h\to 1}\frac{\partial G_i(z_1,z_2,\cdots,z_N,z_h)}{\partial z_j} \quad (4)$$

According to the definition of queue length, the mean queue length is drawn via derived function algorithm as:

$$g_{\mathrm{j1}}(j)=\frac{\lambda_j\sum_{i=1}^{N}(\gamma_{i1}+\gamma_{i2})}{\left(1+\rho_j+\rho_j^2\right)\left(1-\rho_h-\sum_{i=1}^{N}\rho_i\right)} \quad (5)$$

and the mean queue length at the key station is:

$$g_{\mathrm{ih}}(h)=\lambda_h(\gamma_{i1}+\gamma_{i2})+\frac{\lambda_h\rho_i\sum_{i=1}^{N}(\gamma_{i1}+\gamma_{i2})}{1-\rho_h-\sum_{i=1}^{N}\rho_i} \quad (6)$$

2.5 The Analysis of the Mean Inquiry Cycle

Based on the operational mechanism of the system and the theory of queuing, the mean inquiry cycle can be derived as:

$$E[\theta_i]=\frac{\sum_{j=1}^{N}\left(\gamma_{j1}+\gamma_{j2}\right)}{1-\rho_h-\sum_{j=1}^{N}\rho_j} \quad (7)$$

3 Theoretical Calculation, Simulated Experiment and Analysis

In this section, numerical results are presented for the new scheduling algorithm and a comparison is made with traditional RR scheduling scheme of load balance of web cluster servers. Both theoretical and simulated models are set in identical symmetrical conditions under which each station's information packets arrival follows the Poisson distribution with λ as the parameter, and all the station parameters follow the distribution of identical laws. The channel speed is 100Mbps. The time slot width is taken as 15$^{\mu s}$. With $N = 5$, $\beta_h = \beta_i = 10$, $\lambda_h = \lambda_i = \lambda$, $\gamma_{i1} = \gamma_{i2} = 2$, the system meets the stable condition of $\sum_{i=1}^{N} \rho_i + \rho_h < 1$ in which M stands for the number of service times of the controlled gated service at common stations. Analyses are made on the system feature:

1) Fig. 1 and 2 show the effect of the mean queue length of the new scheduling scheme. Under the system stability condition of $\sum_{i=1}^{N} \rho_i + \rho_h < 1$, $0 < \rho_i < 1$, when the gated service at common stations becomes more frequent, the mean queue length diminishes . One or two times gated policy services exert influences on the mean queue length at the common stations. In case of the change with loading, stability and fairness can be maintained via dynamic adjustment of gated service times at common stations.

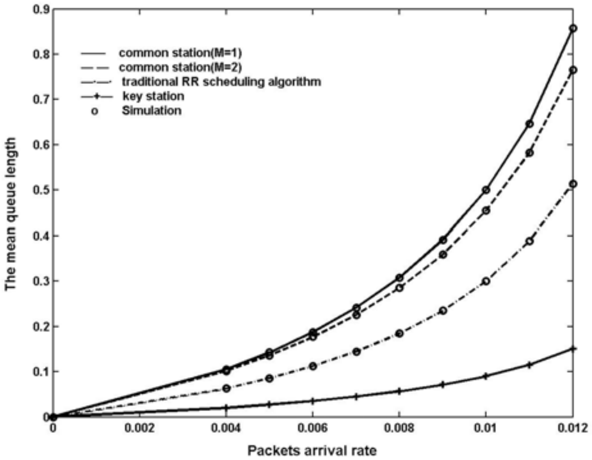


Fig. 1. The mean queue length

2) The change of the mean queue length along with the arrival rate of information packets is illustrated in Fig.1. Compared with the traditional RR scheme, the mean queue length of key station has been markedly less. So, efficiency of the key station is enhanced.

3) Fig.2 shows the results of the comparison between the mean queue length of the key station and the common stations. As indicated, the mean queue length of key station and that of the common stations are clearly differentiated. Even when the arrival rate is high, the mean queue length of key station is smaller but more stable than that of the common stations. Meanwhile, when the gated policy service at the common stations increases, it still maintains a smaller value. The priority of the key station is thus preserved.

4) The results from the calculations via Formulae (6) and (7) and the simulated experiments (Fig.2 and 3) all reveal that in the models of scheme, the change in the

times of service will not influence the mean queue length of key station and the system's mean cyclic time of the model. Compared with the traditional RR scheme, the system's mean cyclic time of the new scheduling algorithm becomes longer.

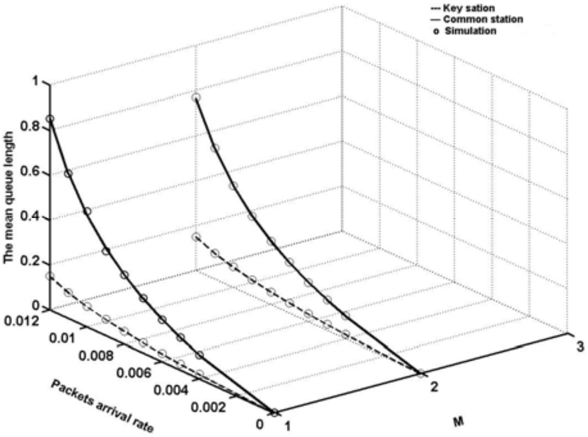


Fig. 2. A comparison between the mean queue length of the key station and the common stations

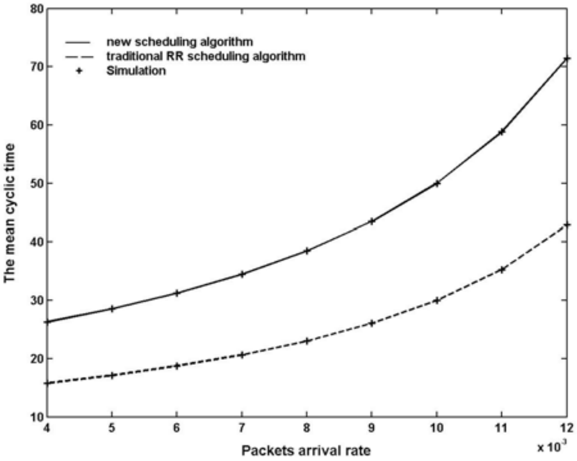


Fig. 3. The mean cyclic time of the system

4 Conclusion

In this paper a new scheduling algorithm of load balance of web cluster servers is proposed. The scheduling scheme classifies the system station by priority at two levels in order to embody the advantages of exhaustive and gated policy service, and it enhances the fairness and flexibility of the service system via dynamic adjustment of gated service times from one to twice at common stations along with the change in network loading. As is shown, the findings from theoretical analysis correspond well with those from simulated experiments.

Acknowledgments. The authors would like to give their sincere thanks to the financial support by the Nation Natural Science Foundation of China (No.F0424104 and No.60362001).

References

1. Cardellini, V., Colajanni, M., Yu, P.S.: Dynamic load balancing on Web-server systems. IEEE Internet Computing 3, 28–39 (1999)
2. Hui, C.-C., Chanson, S.T.: Improved Strategies for Dynamic Load Balancing. IEEE Concurrency (1999)
3. Kant, K., Mohapatra, P.: Scalable Internet Servers: issues and challenges, performance and architecture of Web Servers (PAWS). In: ACM SIGMETRICS 2000 (2000)
4. Damani, O.P., Chung, P.E., Huang, Y., Kintala, C., Wang, Y.: ONE-IP: techniques for hosting a service on a cluster of machines. Journal of Computer Networks and ISDN Systems 29, 1019–1027 (1997)
5. Kant, K., Mohapatra, P.: Scalable Internet Servers: issues and challenges, performance and architecture of Web Servers (PAWS). In: ACM SIGMETRICS 2000 (2000)
6. Iyengar, A., Challenger, J., Dias, D., Dantzig, P.: High Performance Web server design technique. IEEE Internet Computing 4, 17–26 (2000)
7. Menasce, D.A.: Trade-offs in designing Web clusters. IEEE Internet Comput. 6, 76–80 (2002)
8. Baek, S., Rim, H., Kim, S.: Socket-based RR scheduling scheme for tightly coupled clusters providing single-name images. Journal of Systems Architecture 50(6), 299–308 (2004)
9. Dongfeng, Z., Sumin, Z.: Message Waiting Time Analysis for a Polling System with Gated Service. Journal of China Institute of Communications 15, 18–23 (1994)
10. Dongfeng, Z., Sumin, Z.: Analysis of a Polling Model with Exhaustive Service. Acta. Electronica Sinica 22, 102–107 (1994)
11. Wang, Z., Yu, H., Song, Y., Sun, Y.: Characteristics of Mean Period of M1+M2/G/1 Polling System under Mixed Service. Journal of China Institute of Communications 23, 8–18 (2002)

Nature of Chinese Thesaurus Automatic Construction and Its Study of Major Technologies in Digital Libraries

Wen Zeng and Huilin Wang

Institute of Scientific and Technical Information of China,
100038 Beijing, China
{WenZENG,HuilinWANG}@istic.ac.cn

Abstract. Modern information network such as digital library contains much more data than ever before. These data are globally distributed, become accessible to huge, heterogeneous users easily. On the other hand, the enormous amount of information requires powerful tools for the user to find the relevant data. One such tool is thesaurus. The thesaurus as an ontology is playing an increasingly important role in knowledge management and information retrieval of digital library. The paper reconsiders the nature of thesaurus from the view of ontology. It proposes the approach and ideas of Chinese thesaurus automatic construction. Some experiment results and future work are also described in the paper.

Keywords: Ontology, Chinese Thesaurus, Automatic Construction, Digital Library.

1 Introduction

The emergence of a large number of information resources has accumulated a wealth of digital information. The traditional internet technology is only a connection to network resources and does not take into account the structural organization of network resources, resulting in a variety of knowledge and thousands of storage media disorderly, fragmented distribution. How to quickly find information from vast amounts of digital information resources that user needs is an unsolved problem. In order to solve this search problem, the core problem lies in finding an orderly method of organization of information resources. Ontology building is an important element in the information network system. In fact, the building of ontology is essentially to achieve effective organization for knowledge and information.

This paper argues that thesaurus is as a tool for information organization. Digital libraries as a kind of information network, thesaurus is one of important ontology under such information network system. Thesaurus construction is of an important and significance work for the service of digital library. It can achieve efficient organization and utilization of knowledge in the information network. The paper is organized as follows. Section 2 gives the relevant work background on ontology and thesaurus construction. In section 3, we define the nature of thesaurus. In section 4, we introduce the approaches of Chinese thesaurus construction in more details. Furthermore, we provide an overview of the prototype system. The progress of our

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 103–110, 2010.

present research work is described in Section 5.The paper concludes with a brief summary and an outlook on future work in Section 6.

2 Related Work

Ontology is an essential part of many applications. Because of supported by ontology, both the user and the system can communicate with each other using a common understanding of a domain. Although ontology has been proposed as an important and natural means of representing real-world knowledge, most ontology construction are not performed either systematically or automatically. Automatic ontology construction is a difficult task owing to the lack of a structured knowledge base or domain thesaurus. So thesaurus construction is an ontology construction in a sense, it is an important and meaningful task for ontology and information organization or utilization.

At present, the various approaches of thesaurus construction have been recently presented [1-6]. A thesaurus is a controlled vocabulary that shows relations (e.g. semantic) between terms, which can aid searchers in finding related terms to expand queries. Many approaches are reported on the automatic construction of thesaurus. Most approaches rely on the similarity among terms and assume statistical models of the terms as observed in texts to extract the similarity. There are a number of similarity measures available in the literature (McGill, 1979; Fox, 1986; Crouch, 1990; Saiton et al, 1975). In fact, almost all popular traditional thesaurus construction methods are based on NLP. Traditional methods such as co-occurrence analysis, n-gram analysis and tf-idf (term frequency-inverse document frequency) weighting can be used for this purpose. The concept space approach for automatic thesaurus construction was developed by Chen, Ng, Martinez, and Schatz. A concept space is defined as a network of terms and weighted associations which can represent concepts (terms) and their associations for the underlying information space which represents the documents in the database. Thesaurus automatic construction of domestic research can be divided into area thesaurus and comprehensive thesaurus. Comprehensive thesaurus is such as Chinese Thesaurus, E-government Thesaurus and so on. These thesauri are manually finished by experts in the field. For the automatic construction technology, it is mainly tracking the abroad technology research. Many studies are no sense of a complete system research and development work.

3 Nature of Thesaurus

The first definition on ontology is from the knowledge engineering, excepting philosophy, according to Corcho et al. Both the thesaurus and the ontology consist of terms and the relationships between terms. There are many different definitions of thesaurus, varying from quite modest definitions that focus on the relations between words without stating which kinds of relations that are meant, to such definitions that state more exactly which relations that are concerned. A modest definition is presented by Schutze and Pedersen: “We define a thesaurus as simply a mapping from words to other closely related words”. Furthermore, thesaurus for a specific field usually includes a relatively complete set of terms in that field collected and organized by domain experts and knowledge organization professionals, and these terms can be used as concept candidates of domain ontology. As ontology, thesaurus defines a set of representation terms or concept. Inter-relationships among these concepts describe a semantic web. Thesaurus is an important indexing, retrieval and navigation tools in the information resources management. Thesaurus contains all the subject words and non-subject words according to a certain order in a field. Non-subject words are as entry words. Subject word is a concept in which the preferred term for indexing instead of words is an importer of search terms. It makes the concept that has the same meaning can be used to express a word to solve a conceptual problem corresponding to a variety of vocabulary. Each Subject entries composed of a variety of reference matters. Semantic relationship between the thesaurus items is expressed by these references, and references mainly consist of equivalence, hierarchical and associative relationships between the subject words and so on. These relationships can also be used as relationships candidates in the domain ontology.

Most of the structure of the thesaurus is achieved by means of relationships between terms. Among the possible types of relationships, three of them are used for the thesaurus: equivalence, hierarchical and associative relationships. Equivalence relationships are defined between non preferred and preferred terms to denote terms describing the same concept. Equivalence relationships are often used more loosely, including alternative spellings of the same term, upward posting, common and scientific or formal names etc. Equivalence relationships are distinguished into three types: relationships between a non-preferred term and the equivalent preferred term; relationships between a non preferred term and a set of equivalent preferred terms; relationships between a non preferred term and its (preferred terms) syntax components. The hierarchical relationship relates descriptors to its super and subordinate terms (broader terms, narrower terms). Super ordinate terms can be generalizations or subsumptions of descriptors and subordinate terms can be specializations or parts or instances of descriptors. The hierarchical relation is transitive and anti-symmetric and therefore forms a directed acyclic graph. Associative relationships are used to connect semantically related preferred terms which are not hierarchically related. The associative relationship relates descriptors to other descriptors which are not hierarchically related, e.g., if they are opposites. The associative relationship is symmetric.

4 Basic Approaches and Ideas of Chinese Thesaurus Construction

4.1 Approaches of Thesaurus Construction

There are two kinds of thesaurus construction approaches. One is manual thesaurus construction, the other is automatic thesaurus construction. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that affect the structure of the thesaurus. There is a need for methods of automatically construct thesaurus, which besides from the improvements in time and cost aspects can result in more objective thesaurus that are easier to update.

Most of the automatic construction approaches rely on the statistical evidence, such as collocation or co-occurrence, to associate document terms (Park, Han, & Choi, 1995; Srinivasan, 1992). Thesaurus constructed in this way are often called co-occurrence thesaurus. The associations in these thesaurus, although not necessarily as strong in semantics as those in handcrafted ones, often encode the structure of the knowledge underlying the free-text database, a form of knowledge not easily detected or maintained manually. Thus, constructing such a thesaurus is highly related to a knowledge discovery or text mining process for unstructured documents. The concept space approach for automatic thesaurus generation was developed by Chen, Ng, Martinez, and Schatz. A concept space is defined as a network of terms and weighted associations which can represent concepts (terms) and their associations for the underlying information space which represent the documents in the database.

4.2 Basic Ideas and a Proposed System Framework of Chinese Thesaurus Automatic Construction

Institute of Scientific & Technical Information of China had manually finished “Chinese Thesaurus” construction in 1975, and gotten National Science and Technology Progress Award. In the current cross-language, distributed information network, thesaurus is still the effective tool of information organization, retrieval and use. With the development of modern information network such as digital library, Chinese Thesaurus demands to be improved, revised and even reconstructed, especially in some new research area. But thesaurus construction and maintenance must take automatic construction approach, and can not take manual construction approach as before. Based on the demand, we propose our ideas and a system framework of automatic thesaurus construction. The ideas and system framework will be applied to the National Science and Technology Library (NSTL).The Institute of Scientific & Technical Information of China is one of the members of the NSTL, and it is an only authority on the scientific and technical literature collections. Its network service system has put into operation and use in 2000.NSTL is developing towards to the national engineering and technical digital library. Our research work will be based on NSTL and provide related services and technologies. Meanwhile, our research work can be validation and improvement during the process of service.

Our approach will apply machine learning technology Combined with cognitive science, for example data mining and human psychological, to construction thesaurus. The primary system framework of thesaurus construction includes key four processes, namely Document pre-processing, Concept clustering and choosing words, Co-occurrence analysis and Building semantic relation among terms, which are described below.

5 Our Present Study Work

Most of the automatic construction methods can be applied to Chinese documents. However, unlike English, Chinese text has no delimiters between words. Although any number of white spaces can be inserted between characters in written text for better layout, but does not mean that words are segmented by these spaces. Identifying words in a Chinese text is somewhat like identifying multiword phrases in an English document, because both have no lexical delimiters. And the structure of Chinese grammar is more complex than English. Thus, Chinese thesaurus construction is a daunting work. An ideal way would apply natural language processing (NLP) techniques to extract topic-relevant noun phrases as terms for association. However, ideal NLP rely on ideal resources such as dictionaries and corpora for reliable word segmentation and part-of-speech tagging. So, we believe that combining with dictionary or presented thesaurus to improve the quality of choosing words may be a good approach. Meanwhile, Using NLP technology and Chinese grammar knowledge is also important for the construction of Chinese Thesaurus.

At present, we are developing prototype system of Chinese thesaurus automatic construction. We proposed the method of extract Chinese term from text resources, started and finished some experiments. Some details of experiment are as follows.

5.1 Special Domain Documents and Dictionaries Construction

The collection of Special domain document is the basis of thesaurus construction. In order to get high quality domain document, we choose and download latest domain document from NSTL network database and latest web pages from Google. We have built some domain dictionaries, for example computer dictionary, EI term dictionary (EI is one of the international authoritative search institution,we have finished Chinese localization on its term database).

5.2 Document Pre-processing

To construct the domain thesaurus, we convert the document and web pages downloaded into TXT style, and build text databases. Meanwhile, we deal with these Chinese documents, and realize Chinese words cut and POS tagging by Syntax Analysis system that has been developed by our lab.

5.3 Concept Clustering and Choosing Words

In the stage of pre-processing, we achieve the preliminary classification, as well as speech recognition of terminology and concepts. In order to achieve the ultimate extraction of thesaurus terms and the semantic relations between words, further refinement and clustering is necessary. We can use the title and key words as the text clustering basis to achieve further processing, to reduce and improve the size and quality of subsequent extraction. To the terminology extraction, we can select the title, abstract and body as terms source of natural language vocabulary, including nouns, verbs to thesaurus automatic construction. Then, we get meaning Chinese words, phrase and words by choosing noun and matching dictionaries from these results. The parts of results are as follows:

Table 1. The parts of results choosing words

Chinese word	frequency
科技文献翻译	5
概率空间	5
子任务	5
便携设备	5
交通运输	6
模糊语义模式	2

5.4 Co-occurrence Analysis

In our study, the method of co-occurrence analysis is also as a kind of classification method, aim to classification the web pages into a relative correct category.We calculate the similarity values between web documents and sample document.We adopt the VSM model.

$$Document = D(t_1, t_2, ..., t_k) \tag{1}$$

When we give a wight value w_k,then

$$Document = D(t_1, w_1; t_2, w_2; ..., t_n, w_n) \tag{2}$$

$$w_{i,d} = TF_{t,d} \times IDF_t \tag{3}$$

Given document $d_i = (w_{i1}, w_{i2}, ..., w_{in})^T$ and web document: $d_p = (w_{p1}, w_{p2}, ..., w_{pn})^T$ .So their similarity is:

$$Similarity(d_i, d_p) = \frac{\sum_k^n w_{ik} \times w_{pk}}{\sqrt{\sum_{k=1}^n w_{ik}^2} \times \sqrt{\sum_{k=1}^n w_{pk}^2}} \tag{4}$$

We give the threshold value,if the similarity is beyond the theshold,the web page is accepted,otherwise discarded.On the other hand,we also get the kewords by co-occurance analysis as described in Section Ⅳ. We get some experiment results from 1000 documents of maching translation field that have been processed. The part of results are as follows:

Table 2. The part results of co-occurance analysis

rank	word	co-occurrence frequency
1	计算机 (computer)	332
2	机器翻译 (machine translation)	171
3	自然语言处理 (natural language processing)	192
4	人工智能 (artificial intelligence)	144
5	翻译方法 (translation method)	75
6	句法分析 (syntactic analysis)	21

5.5 Building Semantic Relation Among Terms

How to use computer technology to automatically identify semantic relationships between words is the focus of our future research work. It is also one of the difficulties and core problems to achieve Chinese thesaurus automatic construction. At first, we utilize the method that word similarity computing based on How-net. But How-net has only included about 1500 sememes, it can not cover all terms in the science & technology or others areas. For example, we can get the similarity value between computer and software: Similarity(computer,software)=0.444444.But we cannot get the similarity value between machine translation and computer because of lacking the sememe of machine translation in How-net. So we can not certain the relationships between words by How-net. We believe that pattern recognition and data mining technology are effective technology methods to establish semantic relations between words, and combined with human bionics to build words network that contains semantic relations, so-called concept of space is necessary. The concept of space contains any concentration of the relationships between the two terms, this correlation relationship is asymmetrical. Use of concept space approach, we can consider each word in the thesaurus as a model similar to the concept of human space network. The semantics of each term is a sub-network corresponds to entire terminology semantic concept space network of Chinese thesaurus. In this concept of space, we hope to build a relatively good semantic relationship between words through continuous training process. Now, we are doing this work.

6 Conclusion and Future Work

Our study work is only a beginning. In spite of we have accumulated some work basis and experiment results. But we believe that the difficulty of our work is the ability of understanding and analysis to Chinese words, phrases and sentences. It is foundation of distinguishing and building relation among terms in the thesaurus. This will be an emphasis in our future work. In this paper, we think that thesaurus is the ontology in nature. Thesaurus construction is the ontology construction in some sense. We describe the ideas and approach of thesaurus automatic construction. Our ideas and approach are inspired by our future application projects—NSTL. At last, we will implement the ideas, and apply these approaches in the real digital libraries environment. In summary, we believe that Chinese thesaurus automatic construction will be a promising approach to improve the indexing and retrieval process in large information networks such as digital libraries.

Acknowledgments. The project of this paper is supported by the China Postdoctoral Science Foundation funded project (Program No: 20100470392).It is also supported by the State science and technology support program(Program No: 2006BAH03B02).

References

1. Shimodaira, C., Shimodaira, H., Kunifuji, S.: A Divergent-Style Learning Support Tool for English Learners Using a Thesaurus Diagram. Springer, Heidelberg (2006)
2. Alan Wang, G., Chen, H.: A Multi-layer Naive Bayes Model for Approximate Identity Matching. Springer, Heidelberg (2006)

3. Cole, C., Leide, J.: Investigating the Anomalous States of Knowledge Hypothesis in a Real-Life Problem Situation: A Study of History and Psychology Undergraduates Seeking Information for a Course Essay. Journal of the American Society for Information Science and Technology 56(14), 1544–1554 (2006)
4. Yang, M.C., Wood, W.H.: Design information retrieval: a thesaurus-based approach. Springer, Berlin (2005)
5. Ng, C.Y., Lee, J.: Efficient Algorithms for Concept Space Construction. HKU CSIS Tech Report TR-2002-08
6. Bechhofer, S., Goble, C.: Thesaurus construction through knowledge representation. Data & Knowledge Engineering 37(1), 25–45 (2001)

Research on the Resource Monitoring Model Under Cloud Computing Environment

Junwei Ge[1], Bo Zhang[2], and Yiqiu Fang[2]

[1] College of software, Chongqing University of Posts and Telecommunications
400065 Chongqing, China
[2] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications 400065 Chongqing, China
gejw@cqupt.edu.cn, zhanbo1005@163.com, fangyq@cqupt.edu.cn

Abstract. Resource monitoring is an important part of resource management under the cloud computing environment, which provides a better reference for resource allocation, task scheduling and load balancing. Because of the commercial applications target of billing the user for the use of resources , the high virtualization, scalability and transparency of the cloud computing environment's resources, the existing resource monitoring methods of both distributed computing and grid computing can not satisfy the cloud computing environment completely. So, according to the characteristics of cloud computing platforms, we present a novel resource monitoring model appropriately adapted to cloud computing environment, which combines VMM (Virtual Machine Monitor) and the C/C++ called by Java to obtain the information of the resource status. Both theoretical analysis and experiments results show that the model can be used to collect resource monitoring information on nodes and VM (virtual machine), which not only meets the requirements of cloud computing platform features but also has a good property of effectiveness.

Keywords: Cloud computing; virtualization; scalability; virtual machine monitor; resource monitoring model.

1 Introduction

Cloud Computing, an internet-based supercomputing model, was proposed in 2007, which was developed based on the parallel computing, distributed computing and grid computing or which is the commercial implementation of these science concepts [1].

Cloud computing is an innovative IT business model, using the virtualization technology. Cloud computing service providers make the large-scale network servers form the large-scale virtual resource pool, the user only need to configure the low-profile network access devices to access and to use these resources required, which greatly reduce the user's hardware and software purchasing costs. In the cloud computing environment, all of the applications often run in the remote distributed system instead of on a single computer or server. Through using the virtualized and extensible technology, we abstract the hardware and software into the dynamically

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 111–118, 2010.

extensible and configurable resources, and provide to the user through the external service. The creation, publishing, implementation and management of application could all run in the cloud platforms. User only need to bill according to the amount of resources and the scale of application service.

Cluster node is loosely coupled in Cloud computing environment, in order to provide high-quality application services, we have to solve resource management problems, and resource monitoring is also an important component of resource management. In the cloud computing environment, a variety of resources use virtualization technology. Most of the characteristics of resources are hidden, but cloud computing application systems also need to help users find the right resources based on their demand for resources information, such as processor computing speed, memory size, size of hard disk storage space available etc. If a node in the cluster or the virtual machine running in a node has problems, then it requires a system to provide resources for state information, to discover and solve the fault. These are totally dependent on resource monitoring. At the same time, resource monitoring data can be used for resource discovery and allocation; task scheduling and load balancing, according to the features that the user should pay for cloud computing services, the monitoring information play an important reference role in billing the use of resources. Therefore, the research of resource monitoring in cloud computing environment will boost the development of cloud computing.

2 Research Basis

Currently, the research of resource monitoring in the cloud computing is little involved. Through the analysis of cloud computing architecture, we can see the difference among cloud computing, distributed computing and grid computing. The cloud computing architecture is shown in Fig.1.

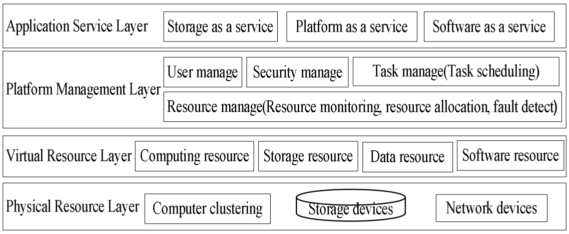


Fig. 1. Cloud computing architecture

(1) Physical Resource Layer: cloud computing services providers need to link the enormous amount of computer clusters and storage devices, through the network equipment at a reasonable topology structure to build the cloud computing data centers.

(2) Virtual Resources Layer: using virtualization technology, the physical layer hardware resources build a virtual resource pool which is allocated according to need and is high degree of sharing.

(3) Platform Management Layer: platform management includes resource management, task management, user management and security management. Dynamic deployment of resources is that the computing resources, storage resources can be quickly allocated to the user. Resource monitoring is that real-time monitor collects the static and dynamic information of each computer in the cluster node and its virtual machines running on it in order to provide the basis for resource discovery, resource allocation and task scheduling. Meanwhile, we should also monitor the cost of resource caused by user tasks. Task scheduling, is that we use efficient scheduling method to achieve load balancing. User management refers to manage the account information of cloud computing services. Security management mainly ensures the security data in the system.

(4) Application Service Layer: Besides using computing services and storage services provided by cloud computing service providers, you can enjoy the platform services, online application services and software services.

Currently resource monitoring has more researched in distributed computing and grid computing. For example DRMonitor[2], Ganglia [3], NWS (Network Weather Service) [4], MDS [5], etc, which have played an important role in the distributed systems and grid system. However, if these are used directly in cloud computing environment, there will be inapposite. On the one hand, the resources is highly virtualization and scalability in cloud computing environment, and the cloud computing also provide services as IaaS, PaaS, SaaS services at different layers. So we should need to monitor not only the physical server resources but also the virtual machine running on it. On the other hand, cloud computing model is for business computing, which needs reasonable billing for users. The granularity of the existing resource monitoring system to monitor information is not fine, which caused the failure of getting the process hierarchy information, and tracking the real-time consumption in CPU, memory and storage resources in the process of implementation of the users. Therefore, considering the characteristics of cloud computing itself, the current resources monitoring methods in distributed computing and grid computing can not fully adapt to the use in cloud computing, but they have some references in cloud computing.

3 The Design and Analysis of Resource Monitoring Model

In the cloud computing environment, resource monitoring has adopted two modes:

Active mode: In the cloud computing data center, working-node is installed and configured the resource monitoring components and the VMM (Virtual Machine Monitor) that adopt a strategic to acquire status information about physical server and virtual machine run on it, then take the initiative send its own monitoring information to the master node.

Passive mode: In the cloud computing data center, the master node sends a request to the working-node, then the working-node respond the data to the master node.

We have to use some polling strategy whether to use any of the two models in the above, because the cloud computing environment resources monitoring information are real-time. Currently, the polling strategy of distributed computing and grid computing are both based on using periodic or event-driven types. Periodic type is that the working-nodes will send the information collected by themselves to the monitor periodically; or master node which use its resource monitoring components send the request to the working-nodes which will respond to master node later when collect their information. The event-driven approach [6] is that the working-node will produce a series of events; each event will trigger the corresponding collector to detect the state of resources monitor and compare to the last value. When the change between the two events is bigger than the threshold that has set before, the working-node will send its own monitoring data actively or passively.

Based on the analysis above of the monitoring pattern of the cloud computing environment and the characteristics of its own, we proposed resource monitoring models applied to the cloud computing environment on the basis of the research of the distributed computing and grid computing. The model is shown in Figure 2:

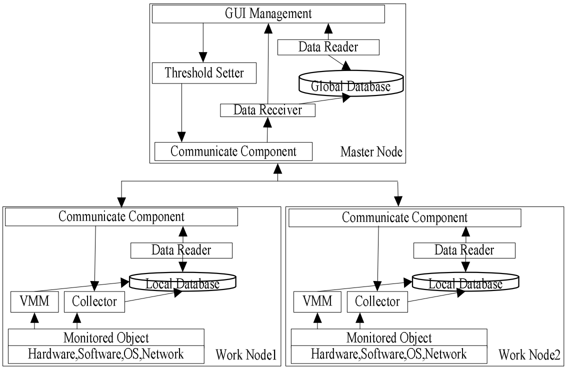


Fig. 2. Resource monitoring model for the cloud computing environment

- Monitored Object: Monitored object of the Cloud computing environment are the hardware resources, software resources, operating system, network resources and so on.
- Collector: periodically collect its own resource data or collect the self-monitoring data by using the threshold strategy for collection.
- VMM (Virtual Machine Monitor): In virtual environment, VMM is responsible for obtaining information about CPU usage, memory usage and network traffic of the virtual machine running on physical server.
- Local Database: Used to store resource data collected by the collector and virtual machine from the work node.
- Data Reader: Used to read the monitoring data from the resource monitor database.
- Communicate Component: the data communication components between Work nodes and master node.
- Data Receiver: from the communication component, master node receives the resources monitoring data from the work node.
- Global Database: to store resource monitoring data which are received by master node in the cloud computing data center.

- GUI Management: system administrator can view monitoring information of the all nodes and set the threshold of the monitored object in the data center through monitor interface.
- Threshold Setter: setting the threshold of monitored object.

In the cloud computing data center, collectors in the each work node are responsible for its own static information: such as the CPU number, type, frequency and other key information, physical memory storage space, virtual memory, storage space, disk storage space, operating system version, IP address, then collected the static information and send it to the Data Receiver of the master node through the communication component which communicate with the master node Data Receiver insert these static information to the Global Database. Static information of each node were written once and for all when they join in the data center, the node static information are automatically deleted when they leaving. The Data Receiver reads the static information stored in the Global Database, and then sends to the GUI Management.

The collector should capture the dynamic information of the node itself periodically or based on the polling strategy of setting the threshold, such as the load average of the node itself, the number of the running tasks and the status of these tasks, CPU usage , occupancy of physical memory, usage of virtual memory, the number of thread of the current running tasks, the running time of each process, the consumption information of CPU and memory of each process, the disk space that has been used and the disk space that are available, the network traffic, network bandwidth and network delays of the current node; VMM is responsible for obtaining the information, such as the CPU, memory usage and network traffic of each virtual machine running on physical server; then write the monitoring data to the Local Database. These data are updated frequently, massive data, and real-time, so there is no meaning to store these data long time. Therefore we should store these data by the queue, the storage space is fixed. When a new data need to store, it will be enqueue, when the queue is full, the monitoring data at the rear of the queue will be dequeue automatically, so the database only store the most recent short-term data. Data Reader reads the monitoring data of the Local Database and sends it to the data receiver in real time through the communication component. These real-time data have no need to store in the global database, so the data receiver sends it to the GUI management directly.

Through the graphical user management interface of the master node monitoring system, the cloud computing platform manager can view the static and dynamic monitoring information of all the nodes, and also can set the threshold for some important monitored objects by the Threshold setter. The threshold setter then passed the threshold value to the collector through the communication component. The collector based on the threshold value; decide to whether to write the monitoring data to the local database. For example, set the CPU utilization threshold value 90%, when the CPU utilization is less than 90%, the collector do not insert the monitoring data to the database, the database retains data before setting the threshold value; when CPU utilization is greater than or equal to 90%, the collector will insert the monitoring data to the database.

4 Implementation Project and Demonstration

Based on the model of resource monitoring system, the data collector and VMM are using C/C++ language, while other components and graphical user interface are written by using Java language. As the C/C++ language can operate the hardware device directly, it is suitable for underlying systems implementation, while the Java language has no link to the OS platform, so it can be run on multiple platforms.

We install Linux system to each node, the collector can obtain the CPU, physical memory, virtual memory, disk space and network equipment data by reading "/ proc" file in system documents and selecting needed data. Use the "ps" command to get the user job information, use the "top" command can also obtain these data and process information, use the "df" command to get the information about the usage of the disk space of the system documents. Use the nload tools which fit in the Linux/Unix systems to monitor the information like network traffic. Monitoring database: according to pre-designed formats, insert the monitoring data to MySQL database in real time. Other parts of components are programmed in Java.

From the analysis and design of the model and implementation of project above, the model has the following characteristics:

(1) Usability. The resource monitoring system based on the model is easy to deploy and has good flexibility. It provides uniform data external access interface and easy to obtain the monitoring data. Meanwhile, the graphical user interface use Java programming, the interface is elegant and easy to use.

(2) Scalability. When the new work nodes join in the cluster of cloud computing environment the monitoring system can directly obtain the static information of the node itself, and insert it into the database. When a work node is leaving, the monitoring system can delete the static information of the node from the global database automatically. Installation and configuration of components in the work nodes is relatively simple, which is helpful for the monitoring system to extend in large-scale cloud computing environments.

(3) Reliability. In the monitoring system, the link that the work node sends the monitoring data to the master node is redundant, so the failure of a single node doesn't affect other nodes.

(4) Alarm function system. The monitoring system provides system alarm function, when a monitored object of a working node has reached a predetermined threshold value, or a working node has problems and can not run, the monitoring system will send alarm information.

(5) Providing a good basis for user's billing pattern. The cloud computing monitoring system based on the model not only can collect static and real-time dynamical information, but also can acquire the process information of the working-node by tracking tasks running status, including the node's list of the current process, each process's CPU resources, memory resources, the process priority, etc, which provide a better basis for the user's billing model.

In the experiment, 3 PCs were used to build cloud computing experimental platform, which includes a computer as the master node, and 2 others as working-nodes. Master node and working-nodes connected together by Gigabit switch. Each node is installed on Linux systems (Ubuntu 9.10) and open source tools Eucalyptus1.6.2 [7], which is as the cloud computing platform, in which we can develop the resources monitoring system based on our model. VMM use Xen HVM3.1.0. Virtual machine of the physical is installed on CentOS4.5 whose memory is 512M and the disk image is 8GB.

The main purpose of this experiment is to obtain the CPU occupied rate, memory capacity and average load of the virtual machine running in the physical machine. The fig. 3 and fig. 4 shows the resource monitoring information of virtual machine1 in node1 and virtual machine1 in node2 separately as the sampling time is 500s, collecting cycle is 1s. From top to bottom is CPU occupancy rate consumed by the virtual machine, CPU occupancy rate consumed by one process of the virtual machine, the average load of the virtual machine system and the memory condition consumed by the virtual machine.

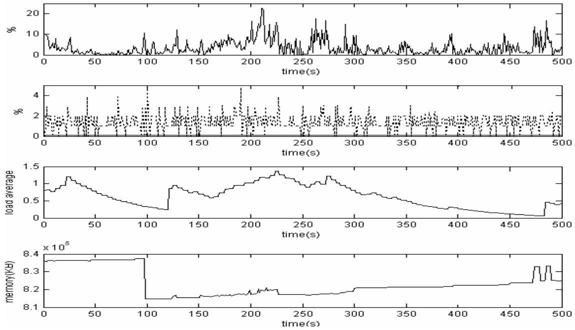


Fig. 3. Monitoring information of virtual machine 1 on workering-node1

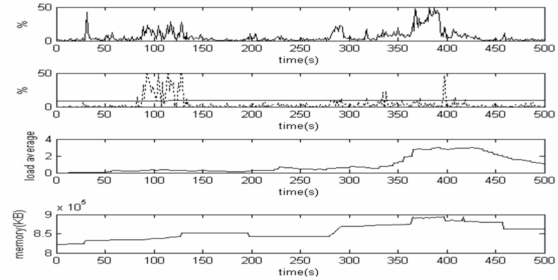


Fig. 4. Monitoring information of virtual machine 1 on workering-node2

Through the above experimental results, shown in fig. 3 and fig. 4, the resource monitoring models proposed in this paper can effectively obtain monitoring information consumed by the virtual machine running in the physical machine. We can also obtain the information consumed by the process through further tracing a process, which can provide the foundation of cloud computing further commercial application.

5 Conclusion and Further Study

Resource monitoring is an essential part for cloud computing platform. In this paper, we proposed a resource monitoring models in cloud computing environment, and analyzed the key technologies in order to achieve this model. We also analyzed the superiority of this model, which is usability, scalability, reliability, system alarms and providing a better reference for user to use the accounting pattern. Finally, the experimental result indicates that the proposed model can obtain monitoring information of the virtual machine effectively. The resource monitoring system based on the model can collect both the real-time working-node's static and dynamical information and virtual machine's information that is running on the node, which can provides a better reference for resource discovery, resource scheduling and load balancing. On the basis of this, the further research work is according to the data obtained by resource monitor, one can design efficient and reliable resource discovery algorithm to meet the user's request to find the necessary resources.

Acknowledgements

The author would like to thank the sponsors of the Science and Technology Research Projects of Chongqing Education Committee under Grant No. KJ090519.

References

1. Guo, B., Wang, P., Chen, G.: Cloud computing model based on MPI. Computer Engering 12(24), 84–86 (2009)
2. Smith, G., Baker, M.: A Flexible Monitoring and Notification System for Distributed Resources. In: International Symposium Parallel and Distributed Computing, pp. 31–38 (2008)
3. Liu, Y., Gao, S.: WSRF-Based Distributed Visualization. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 615–619 (2009)
4. Zhang, J., Figueiredo, R.: Adaptive Predictor Integration for System Performance Prediction. In: Parallel and Distributed Processing Symposium, pp. 1–10 (2007)
5. Diaz, I., Fernandez, G., Martinm, M.: Integrating the common information model with MDS4. In: 9th IEEE/ACM International Conference on Grid Computing, pp. 298–303 (2008)
6. Fang, L., Hang, T., Shu, J.: Study on energy monitoring mechanism for event-driven wireless senstor networks. Sensor and Micro System 27(10), 14–17 (2008)
7. Nurmi, D., Wolski, R., Grzegorczyk, C.: The Eucalyptus Open-Source Cloud-Computing System. In: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 124–131 (2009)

A New Photo-Based Approach for Fast Human Body Shape Modeling

Xiao Hu Liu, Yu Wen Wu*, and Yan Ting Huang

Engineering Computing and Simulation Institute,
Huazhong University of Science and Technology, Wuhan, 430074, P.R. China
xhliu@mail.hust.edu.cn, atyuwen@smail.hust.edu.cn,
hytpauls@163.com

Abstract. In this paper, a new photo-based approach for fast human body shape modeling is proposed. This approach takes the user's front and side digital photos as input, and then quickly generates a semblable body shape model with quadrangle mesh. The approach includes five steps: (1) extract profiles from user's photos and divide body contour into parts, (2) automatically select body part from pre-build body parts database, (3) conduct parametric axial deformation for each part, (4) refine body parts based on FFD, (5) smooth the body mesh. The algorithms are elaborated in detail. Experiments show that this approach can achieve good results very quickly and cheaply.

Keywords: virtual fitting, human body modeling, axial deformation, FFD.

1 Introduction

It was predicted that the cloth online fitting room will be a promising web application. Many efforts have been made to develop virtual cloth try-on systems, for example, Ref. [1-4], among others. In order to make those systems practical, a fast virtual body modeling approach is necessary. Some researchers resort to build a standard generic model and then deform it based on a cluster of body measurements provided by user [5]. While small amount of measurements lead to a poor body model, big amount of measurements make the method unpractical because it is inconvenience for users to do lots of measures and input them one by one. Some researchers proposed a body modeling method from multi-view photos [6], but it often caused distortions and missed some important features.

Photo-based body modeling methods are simple and efficient. However, there is a common fatal flaw in them. The photo profiles are two dimensional curves and it is difficult to represent the complicated three dimensional human body surfaces by these 2D curves. In fact, photo profile should be used to adjust an existed body surface, but not to build a new one.

In order to improve previous photo-based body modeling methods, we propose a hybrid approach. In our approach, the body is divided as five groups, i.e., arm, chest,

[*] Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 119–127, 2010.

waist, stern, and leg groups. A database containing all these parts with various shapes is setup in advance. By matching, deforming, and refining the parts in the library with user's front and side photo profiles, a semblable virtual body can be generated quickly.

2 Overall Description of the Approach

There are five main steps in our approach for fast human body modeling, see Fig. 1. First, user provides two color or black and white digital photos (one front and one side) of his/her own. There is no strict restraint on the resolution of the two photos, but it is recommended that the backgrounds be as simple as possible. The user's body contours are extracted and divided into parts, e.g. arm, chest, waist, etc. automatically. Meanwhile, some key measurements and points of the body are obtained.

In the second step, a set of points on the border edges is sampled for each body part. The Hausdorff distance between a pair point sets is calculated and used as a criterion for automatic selection of the best matching part from the pre-build body parts database. The selected parts are joined together to form a coarse body model.

In the third step, parametric body modification is conducted via axial deformation for every selected part, according to the body measurements in the second step.

In the fourth step, by taking the contours as reference, a local free-form deformation is adopted to refine the body model.

In the last step, in order to get fair body surfaces, we resort to Laplacian method to smooth the body mesh.

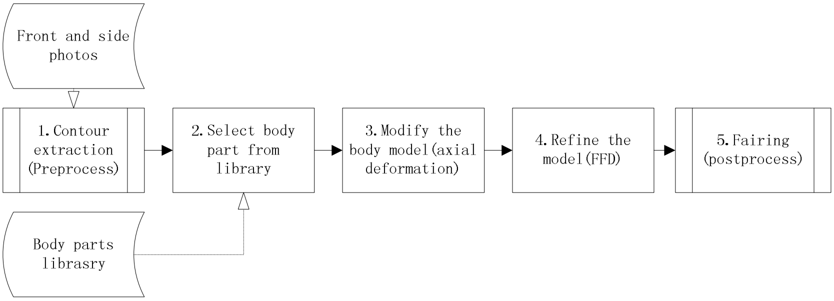

Front and side photos
1. Contour extraction (Preprocess)
2. Select body part from library
3. Modify the body model (axial deformation)
4. Refine the model (FFD)
5. Fairing (postprocess)
Body parts librasry

Fig. 1. Flow chart of the method

3 Details and Algorithms

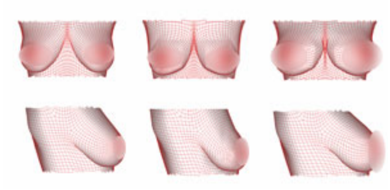
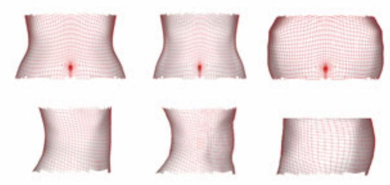
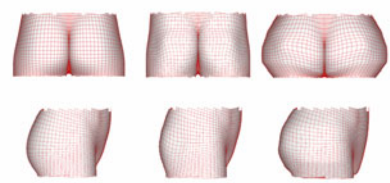
After users provide their front and side digital photos with simple background, the photo contours can be extracted. There are numbers of photo edge detection and contour tracing algorithms, but none of them is perfect. For simplicity, a Sobel operator is adopted to clear the background pixels and extract the contours.

From these photo contours, a group of body measurements can be extracted, such as stature, chest measurement, waist measurement, leg length, and so on. To do this, we use a method similar to that of Hilton et al.'s reported in Ref. [6]. First, some feature points such as calvaria, vola, crotch, armpit, etc, are marked on the photo profile by user. And then these points, along with some empirical equations, are used to divide the body into seven parts: head, arm, neck, chest, waist, stern, and leg. The measurements all the body parts except head and neck are obtained automatically. The head and neck parts are not included in our body model.

3.1 Body Part Selection

Shapes of human body are sometimes totally different. The body measurements can show the size and general features of the human body, however, the actual 3D body shapes cannot be well represented by measurement data. Say two women have the same chest measurement, but the shapes of their chests may be very different. Therefore, we sampled a number of men and women with different body shapes to get various body part models, and used them to establish a body parts database (or BPDB for brevity). Some examples of body part models (meshed by quadrangle) are shown in Table 1. In the body modeling process, each body part is automatically selected from BPDB and joined together to form a whole body. Users are also allowed to choose other body part manually to replace the automatically selected one.

Table 1. Examples of body parts model

Body Part	Meshed model
Chest	
Waist	
Stern	

To select the body part which matches the photo contour best from BPDB automatically, we take a modified Hausdorff distance [7-8] as criterion. Consider two sets of points: $A=\{a_1,a_2,a_3,...,a_m\}$, and $B=\{b_1,b_2,b_3,...,b_n\}$, the modified Hausdorff distance is defined as following

$$h^{1/2}(A,B)=f^{1/2}_{a\in A}\min_{b\in B}\|a-b\| \tag{1}$$

Where $\|a-b\|$ denotes the Euclid distance between points a and b , and $f^{1/2}_{x\in X}g(x)$ means to take the median over set X . In order to evaluate the similarity between a body part model M in BPDB and the user's photo contour group C , two point sets M_f and M_s are extracted by sampling the border of model frontward and sideward, respectively. Also two other point sets C_f , C_s are cut out from C , see Fig. 2. Then the discrepancy can be calculated in terms of the modified Hausdorff distance

$$d(M,C)=h^{1/2}(M_f,C_f)+h^{1/2}(M_s,C_s) \tag{2}$$

In order to preserve translation invariance and scaling invariance, the coordinates of points are normalized in range $\{(x,y)\,|\,x,y\in[0,1]\}$, with reference to its bounding rectangle, see the bottom left in Fig. 2.

Now walk though the BPDB, for each model, calculate its discrepancy with the photo contour group, eventually the one with lowest discrepancy is selected.

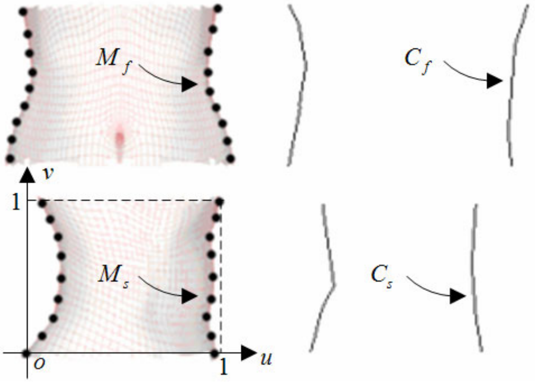


Fig. 2. The four point sets: M_f (*upper left*) and M_s (*lower left*) are extracted by sampling the border of model frontward and sideward; C_f (*upper right*) and C_s (*lower right*) are cut out from the front photo contour and the side photo contour, respectively

3.2 Parametric Axial Deformation

So far we have obtained a rough body model meshed by quadrangle, next we will use the body measurements extracted from the photo contours to stretch the mesh via axial deformation. Axial deformation is an intuitive deformation technique proposed by Lazarus et al [9], which use a 3D axis as reference to build a local frame, and then any deformation applied to the axis can be passed on to the object mesh correspondingly.

Suppose we want to deform a meshed arm portrayed in Fig. 3. Firstly, define an axis L with two control points $\mathbf{C}_1$ and $\mathbf{C}_2$, then for any point $\mathbf{P}$ in the mesh, its projection point $\mathbf{Q}$ on L can be determined as

$$\mathbf{Q}=\frac{(\mathbf{P}-\mathbf{C}_1)\bullet(\mathbf{C}_2-\mathbf{C}_1)}{(\mathbf{C}_2-\mathbf{C}_1)\bullet(\mathbf{C}_2-\mathbf{C}_1)}(\mathbf{C}_2-\mathbf{C}_1)+\mathbf{C}_1 \tag{3}$$

Since we rarely need to twist the mesh along the axis, so point $\mathbf{P}$ can be parameterized with reference to L simply by a scalar r and a vector $\mathbf{n}$

$$\begin{aligned} r&=\frac{(\mathbf{Q}-\mathbf{C}_1)\bullet(\mathbf{C}_2-\mathbf{C}_1)}{(\mathbf{C}_2-\mathbf{C}_1)\bullet(\mathbf{C}_2-\mathbf{C}_1)} \\ \mathbf{n}&=\mathbf{P}-\mathbf{Q} \end{aligned} \tag{4}$$

In this way, when L is stretched to some specified length (e.g., the arm length measurement extracted from photo contours), and the mesh is circumferentially scaled to some specified size (e.g., the arm size measurement extracted from photo contours), the point $\mathbf{P}$ can be relocated by using these two parameters

$$\mathbf{P}=r(\mathbf{C}_2-\mathbf{C}_1)+\mathbf{C}_1+s\mathbf{n} \tag{5}$$

Where s is the scale factor, namely, the ratio of the new arm size to the old size. And if necessary, twisting and anisotropic scaling can also be achieved by adding an angle parameter in the algorithm described above.

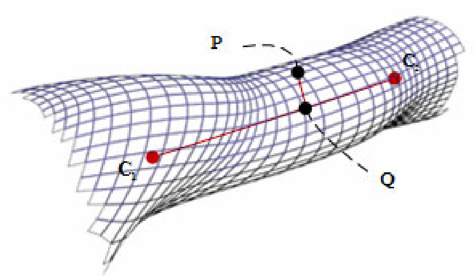


Fig. 3. A meshed arm. Two control points $\mathbf{C}_1$ and $\mathbf{C}_2$ define an axis L. $\mathbf{P}$ is a point on the surface, and $\mathbf{Q}$ is $\mathbf{P}$'s projection on L.

3.3 Silhouette Refinement

To refine the model to match the photo contour, a local free-form deformation method (FFD) is adopted, which is introduced by Sederberg and Parry in Ref. [10]. The key idea of FFD is to embed the object which we wish to deform into a meshed parallelepiped, and then any deformation on the parallelepiped will warp the embedded object accordingly. Take the lattice nodes as control points, any point after deformation can be located in terms of a tensor product of trivariate Bernstein polynomial

$$\mathbf{P}=\sum_{i=0}^{l}\sum_{j=0}^{m}\sum_{k=0}^{n}b_{i,l}(s)b_{j,m}(t)b_{k,n}(u)\mathbf{C}_{ijk} \tag{6}$$

Where l, m, n is the number of slices along the three axes of the local frame on the parallelepiped grid, respectively. $\mathbf{C}_{ijk}$ is the position of control point, and s, t, u $(0 \le s \le 1, 0 \le t \le 1, 0 \le u \le 1)$ is the parametric coordinate of $\mathbf{P}$. $b_{v,n}(x)$ is Bernstein basis polynomial which is defined as

$$b_{v,n}(x) = \binom{n}{v} (1-x)^{n-v} x^{v} \tag{7}$$

First, we divide the body part to be refined and the corresponding front and side photo outlines into K slices. For each layer, we calculate the four offsets, namely, the *offxl*, *offxr*, *offzl* and *offzr*, as illustrated in Fig. 4. Then an axis aligned bounding box (AABB) for the body part is built and meshed by $K \times M \times N$ grid. The displacement of each lattice node can be obtained via linear interpolation of the four offsets in proper layer

$$\begin{aligned} dispx(k,m,n) &= \frac{m}{M} offxl(k) + \frac{M-m}{M} offxr(k) \\ dispz(k,m,n) &= \frac{n}{N} offzl(k) + \frac{N-n}{N} offzr(k) \end{aligned} \tag{8}$$

Where k, m, n $(0 \le k \le K, 0 \le m \le M, 0 \le n \le N)$ denotes the index of the lattice point. Coordinate of point can be obtained by simply adding the displacement to its initial position.

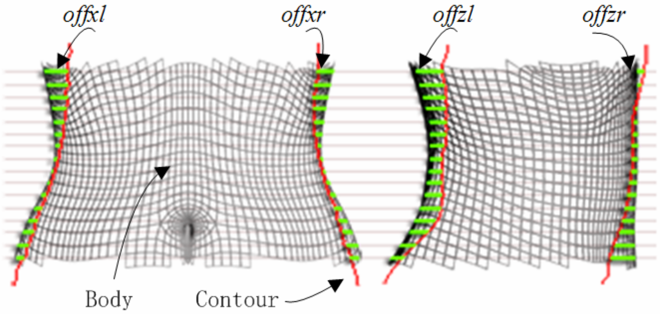


Fig. 4. Slices of the model and contour

In order to simplify the calculation, for any point $\mathbf{P}$, only its eight nearest neighbor nodes in the grid is regarded as control points. Thus equation (6) degenerates to a plain trilinear interpolation

$$\mathbf{P} = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} (1-s_L)^{1-i} s_L^i (1-t_L)^{1-j} t_L^j (1-u_L)^{1-k} u_L^k \mathbf{C}'_{ijk} \tag{9}$$

Where s_L, t_L, u_L $(0 \le s_L \le 1, 0 \le t_L \le 1, 0 \le u_L \le 1)$ is the parametric coordinate of $\mathbf{P}$ with reference to its outer cell. $\mathbf{C}'_{ijk}$ denote the positions of the eight control points as illustrated in Fig. 5.

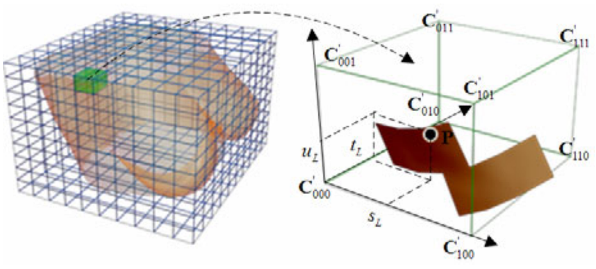


Fig. 5. Parametric coordinate in a cell. $\mathbf{C}'_{ijk}$ denote the eight nearest control points, and (s_L, t_L, u_L) is the parametric coordinate of point $\mathbf{P}$.

4 Applications and Results

We used six groups of photos (each group contains a front photo and a side photo) as experimental cases to test our method. Photos in case 1 to 4 are taken by us, photos in case 5 to 6 are obtained from the Internet [11]. At the present stage, we only built a very small BPDB for the experimental purpose. It takes only three seconds on a desktop PC (Intel Core 2 Duo E6600 CPU, 3GB RAM) to generate the final body model for each case. The results are shown in Fig. 6-7, in which, the left column,

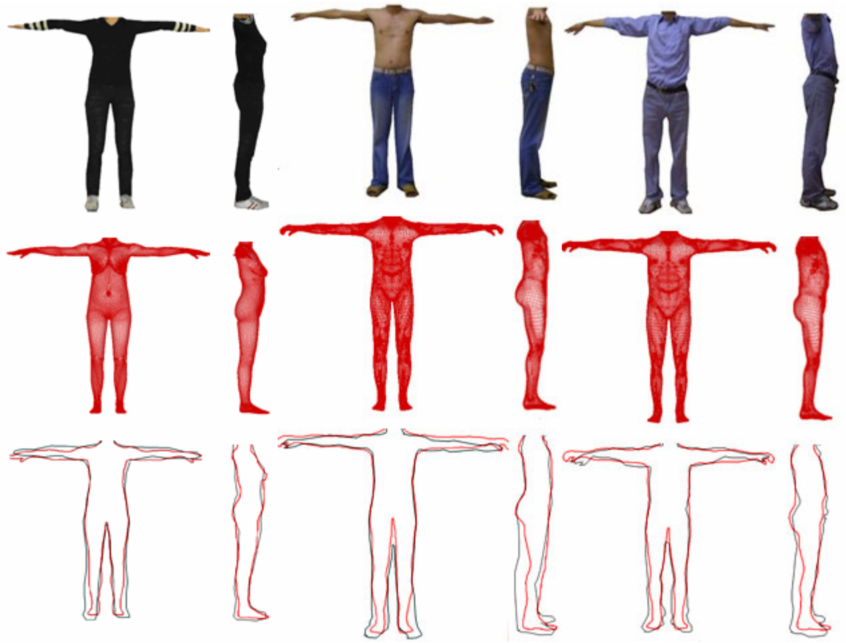


Fig. 6. Case 1 (*left column*), case 2 (*middle column*) and case 3 (*right column*)

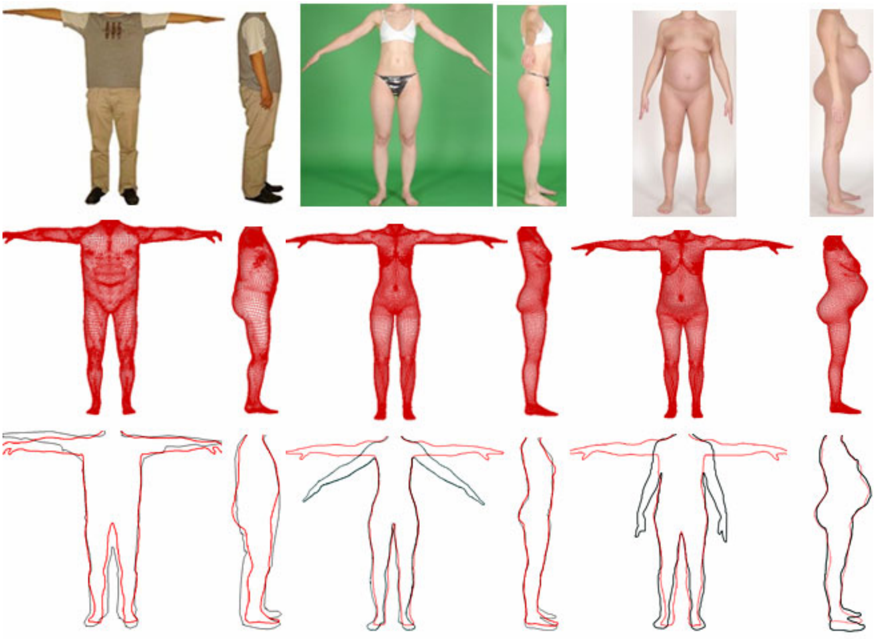


Fig. 7. Case 4 (*left column*), case 5 (*middle column*) and case 6 (*right column*)

middle column and right column show three cases independently, and in each column, the original photos (front and side) are on the top, front and side views of the generated virtual body are listed in the middle, and the contours of photo (shown in black line) and the virtual body (shown in red line) are drawn together in the bottom.

5 Conclusion

We proposed a new photo-based approach for human body modeling that can be used in applications such as online cloth virtual fitting room. It was shown that our approach can quickly generate meshed semblable virtual bodies for users. However, accuracy of our method is fairly influenced by the accuracy of the contours extracted from photos. In order to get better results, sometimes the user is required to do some extra pro-process on the photos. In the refinement step of our method, rough photo contours leads to rugged models, and a Laplacian smoothen method always shrinks the model. In the future work, we will find a way to get more precise and smooth contours from photos with fewer user interventions.

Acknowledgments

The funding support of this work by Natural Science Fund of China (NSFC) (Grant No. 10672066) is greatly appreciated.

References

1. http://www.dressingsim.com
2. http://www.browzwear.com
3. Liu, X.H., Wu, Y.W.: A 3D display system for cloth online virtual fitting room. In: 2009 World Congress on Computer Science and Information Engineering, vol. 7, pp. 14–18 (2009)
4. Liu, X.H., Jiang, C.F., Sze, K.Y., Wang, C.: Online Cloth Virtual Fitting Room Based on a Local Cluster. In: 2009 International Conference on New Trends in Information and Service Science, pp. 139–144 (2009)
5. Protopsaltou, D., Luible, C., Arevalo, M., et al.: A body and garment creation method for an internet based virtual fitting room. In: Proceedings of Graphics International Conference, pp. 105–122. Springer, Heidelberg (2002)
6. Hilton, A., Beresford, D., Gentils, T., et al.: Whole-body modeling of people from multiview images to populate virtual worlds. In: The Visual Computer, vol. 18, pp. 411–436. Springer, Heidelberg (2002)
7. Zhang, D.S., Lu, G.J.: Review of shape representation and description techniques. Pattern Recognition 37, 1–19 (2004)
8. Rucklidge, W.J.: Efficient locating objects using Hausdorff distance. Int. J. Comput. Vision 24(3), 251–270 (1997)
9. Lazarus, F., Coquillart, S., Jancene, P.: Axial deformations: an intuitive deformation technique. Computer Aided Design 26(8), 607–613 (1994)
10. Sederberg, T.W., Parry, S.R.: Free-form deformations of solid geometric models. In: Siggraph 1986 Proceedings, vol. 20, pp. 151–160 (1986)
11. http://bbs.hxsd.com/showthread.php? t=7896134 & page=7

A Multi-Criteria Analysis Approach for the Evaluation and Selection of Electronic Market in Electronic Business in Small and Medium Sized Enterprises

Xiaoxia Duan, Hepu Deng, and Brian Corbitt

School of Business Information Technology and Logistics
RMIT University
Melbourne, Victoria 3001, Australia
{xiaoxia.duan,hepu.deng,brian.corbitt}@rmit.edu.au

Abstract. This paper presents a multi-criteria analysis approach for effectively evaluating and selecting the most appropriate electronic market (e-market) in electronic business by extending the technique for order preference by similarity to ideal solution (TOPSIS). The subjective assessments of the decision maker in the e-market evaluation and selection process are represented by linguistic variables approximated by fuzzy numbers. The geometric centre based defuzzification method is used for transforming the weighting fuzzy performance matrix into the crisp performance matrix on which the TOPSIS is applied for calculating the overall performance of individual e-markets across all the selection criteria and their associated sub-criteria. An example is presented for demonstrating the applicability of the approach for solving the e-market evaluation and selection problem.

Keywords: Multi-criteria decision analysis, Electronic market.

1 Introduction

Electronic market (e-market) is a virtual marketplace in which buyers and sellers are brought together in one central market for exchanging goods, services or information (Dou and Chou, 2002; Grieger, 2003). It has become increasingly popular due to its potential benefits to business, especially to small and medium sized enterprises (SMEs). The existence of various e-markets with their own characteristics, however, complicates the evaluation and selection of specific e-markets in electronic business for SMEs. As a result, evaluating and selecting the most appropriate e-market for SMEs in electronic business becomes a challenging task.

Several approaches are developed for assisting SMEs with their evaluation and selection of e-market in electronic business. For example, Stockdale and Standing (2002) present a content analysis based approach for the selection of e-market. Buyukozkan (2004) develops an index-oriented approach for determining the overall performance of individual e-markets with the use of the fuzzy analytic hierarchical process. Hopkins and Kehoe (2007) propose a matrix-based approach for facilitating the evaluation and selection of e-market while considering the specific requirements

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 128–137, 2010.

of customers. These developments provide SMEs with important means for their evaluation and selection of e-market in their pursuit of electronic business.

Existing approaches, however, are not totally satisfactory due to the inadequacy of handling the subjectiveness and imprecision in the evaluation process and the computational effort required. Furthermore, these approaches have not specifically addressed the nature of individual e-markets and the characteristics of individual organizations. The development of a simple and effective approach capable of addressing the above shortcomings is thus desirable.

This paper presents a multi-criteria analysis approach for effectively evaluating and selecting the most appropriate e-market in electronic business by extending the technique for order preference by similarity to ideal solution (TOPSIS). The subjective assessments of the decision maker in the e-market evaluation and selection process are represented by linguistic variables approximated by fuzzy numbers. The geometric centre based defuzzification method is used for transforming the weighting fuzzy performance matrix into the crisp performance matrix on which TOPSIS is applied for determining the overall performance of individual e-markets across all the selection criteria and their associated sub-criteria. An example is presented for demonstrating the applicability of the approach for solving the e-market evaluation and selection problem.

In what follows, an overview of e-market evaluation and selection is presented, leading to the identification of the selection criteria and their associated sub-criteria. A multi-criteria analysis approach is then presented in Section 3 for evaluating and selecting the most appropriate e-market. An example is given in Section 4, followed by the conclusion of this paper in Section 5.

2 An Overview of E-market Evaluation and Selection

The evaluation and selection of the most appropriate e-market for electronic business in SMEs is complex and challenging. The process of making the selection decision requires the decision maker to simultaneously consider both the nature of individual e-markets available and the specific characteristics of the SME involved. A comprehensive literature review of the specific characteristics of SMEs and the nature of e-market leads to the determination of the selection criteria and their associated sub-criteria in the adoption of e-market in electronic business. Four main criteria are identified including (a) the e-market Capability (b) the e-market Attractiveness (c) the SME’s Capability, and (d) the Electronic Business Environment for the evaluation and selection of e-market. Fig. 1 shows the hierarchical structure of the e-market evaluation and selection problem in SMEs.

E-market capability describes the capacity of individual e-markets based on the deployment of their resources and functionalities for fulfilling the need of individual SMEs in electronic business (Milliou and Petrakis, 2004). It can be measured by three sub-criteria including the market orientation, the market revenue model and the technological competency.

The market orientation of an e-market refers to the specific buyer and seller segments that an e-market aims to serve (Ravichandran et al., 2007). Without a clear focus, the e-market is in the position of selling everything to everyone, which in turn means selling nothing to anybody (Brunn et al., 2002). The revenue model of an e-market determines how an e-market charges the customer on the service it provides. A well-designed revenue model of an e-market helps individual SMEs attract more customers, leading to better performance of the organization in electronic business. The technological competency of an e-market concerns about the design of the technological platform in the e-market. The technological platform should be able to support the development of advanced market-making tools, integrated procurement tools and advanced collaboration tools for e-market. (Stockdale and Standing, 2002).

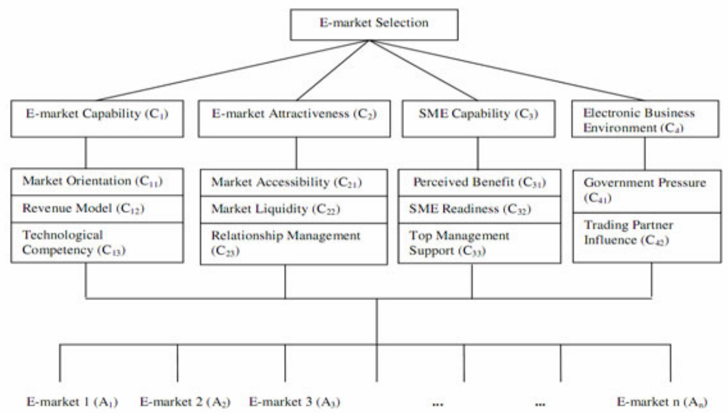

E-market Selection
E-market Capability (C_1)
E-market Attractiveness (C_2)
SME Capability (C_3)
Electronic Business Environment (C_4)
Market Orientation (C_{11})
Revenue Model (C_{12})
Technological Competency (C_{13})
Market Accessibility (C_{21})
Market Liquidity (C_{22})
Relationship Management (C_{23})
Perceived Benefit (C_{31})
SME Readiness (C_{32})
Top Management Support (C_{33})
Government Pressure (C_{41})
Trading Partner Influence (C_{42})
E-market 1 (A_1)
E-market 2 (A_2)
E-market 3 (A_3)
...
...
E-market n (A_n)

Fig. 1. A Hierarchical Structure of E-market Evaluation and Selection Problem

The e-market attractiveness refers to the power of available services, the ability to add values for the customers and the relationship between the e-market and the customer (Buyukozkam, 2004; Standing and Lin, 2007). Without an apparent difference between the supplied values of two e-markets, the customer can easily switch between them. The e-market attractiveness is determined by market accessibility, market liquidity, and relationship management.

The market accessibility of an e-market is related to industry knowledge, market expertise and right product or service determination especially in the right time to create a powerful value proposition towards its target market (Buyukozkam, 2004). The market liquidity of an e-market refers to the volume of transaction conducted. The higher volume of the transaction conducted in an e-market, the more likely an organization in e-market would survive (Brunn et al., 2002). The relationship management relates to the trust and privacy issues in an e-market (Standing and Lin, 2007). The more trust SMEs have on an e-market, the more likely they would select the particular e-market for their e-business.

SME capability can be determined by the perceived benefit, the SME readiness and the top management support. The perceived benefit is related to the benefits such as increasing price transparency, saving operation costs and improving the company's image that SMEs perceive an e-market can bring (Daniel et al., 2004; Standing and Lin, 2007). The SME readiness refers to the level of financial resources and technological resources available in the organization to support the adoption of an e-market (Stockdale and Standing, 2002). The top management support concerns about creating a supportive climate to facilitate the e-market adoption in SMEs (Delone, 1988). It ensures the limited resources and technical expertise to be allocated to support the essential needs of e-market.

The electronic business environment is related to the circumstance in which SMEs conduct their electronic business (Bunker and MacGregor, 2000). The environment that facilitates SMEs' adoption of e-market includes the support from the government and the pressure that the trading partners have placed on individual SMEs.

The support from the government is reflected by the development of appropriate policies and strategies for improving the business environment in assisting SMEs with adoption of the latest technology for their electronic business (Stockdale and Standing, 2002). The pressure from trading partners is another facilitator for SMEs in adopting a specific e-market for their electronic business. SMEs by the nature have little control over the environment. They are more likely to be economically dependent on the government or bigger trading partners for their survival (Bunker and MacGregor, 2000). Pressures from both parties therefore affect the decision of SMEs in adopting an e-market.

To effectively evaluate and select the most appropriate e-market from available e-markets in a given situation, the decision maker in SMEs needs to simultaneously consider the multiple and usually conflicting selection criteria as discussed above. Subjective and imprecise assessments are present in determining the relative importance of selection criteria and assessing the performance of individual e-markets with respect to a specific criterion. To facilitate SMEs' evaluation and selection of the most appropriate e-market in electronic business, the development of a simple and effective approach capable of addressing the above issues is obviously desirable.

3 A Multi-criteria Analysis Approach

Multi-criteria analysis approaches are proven to be effective in tackling problems involving in evaluating and selecting alternatives from a finite number of alternatives with respect to multiple, often conflicting criteria (Deng et al., 2000, Deng and Yeh, 2006). The multi-dimensional nature of the e-market evaluation and selection process justifies the use of the multi-criteria analysis methodology for solving the e-market evaluation and selection problems.

TOPSIS is a popular multi-criteria analysis approach for solving various multi-criteria analysis problems in different areas such as politics, economics, social and management science (Chen and Hwang, 1992). The underlying rationale of this approach is that the most preferred alternative should have the shortest distance from the positive ideal solution and at the same time have the longest distance from the negative ideal solution. The popularity of TOPSIS in addressing various practical problems is due to its simplicity and comprehensibility in concept and efficiency in calculation (Deng et al., 2000).

Subjectiveness and imprecision are always present in e-market evaluation and selection due to the presence of (a) incomplete information (b) conflicting evidence, (c) ambiguous information, and (d) subjective information (Chen and Hwang, 1992; Yeh et al., 2000). To adequately solve the e-market evaluation and selection problem, this section extends the TOPSIS for effectively modeling the subjectiveness and imprecision inherent in the human decision making process with the use of linguistic variables approximated by fuzzy numbers.

A typical e-market evaluation and selection problem can be characterized by (a) the available e-markets for evaluation and selection, denoted as alternatives A_i $(i=1, 2, ..., n)$ and (b) the multiple evaluation and selection criteria C_j $(j = 1, 2, ..., m)$ and their associated sub-criteria C_{jk} $(k = 1, 2, ..., p_j)$ as shown in Fig.1. The e-market evaluation and selection process involves in (a) assessing the performance ratings of each e-market with respect to the selection criteria and sub-criteria as x_{ij} $(i = 1, 2, ..., n, j = 1, 2, ..., m)$, (b) determining the relative importance of the criteria as criteria weights $W = (w_1, w_2, ..., w_j)$ and their associated sub-criteria as sub-criteria weights $W_j = (w_{j1}, w_{j2}, ..., w_{jk})$, and (c) aggregating the performance ratings and criteria weights for determining the overall performance of individual e-markets on which the selection decision can be made.

To adequately model the subjectiveness and imprecision of the e-market evaluation and selection process, linguistic variables approximated by triangular fuzzy numbers are used for representing the decision maker's subjective assessments of the criteria weightings and alternative performance ratings. Triangular fuzzy numbers is usually denoted as (a, b, c) in which b is used to represent the most possible assessment value, and a and c are used to represent the lower and upper bounds used to reflect the fuzziness of the assessment (Deng et al., 2000). Table 1 shows the approximate distribution of the linguistic variables Performance and Importance for measuring the alternative performance rating and criteria weightings respectively in the e-market evaluation and selection process.

Table 1. Linguistic Variables and Their Corresponding Triangular Fuzzy Numbers

Performance		**Importance**	
Linguistic Variable	***Fuzzy Numbers***	***Linguistic Variable***	***Fuzzy Numbers***
Very Poor (VP)	(0.0, 0.0, 0.3)	Very Low (VL)	(0.0, 0.0, 0.3)
Poor (P)	(0.1, 0.3, 0.5)	Low (L)	(0.1, 0.3, 0.5)
Fair (F)	(0.3, 0.5, 0.7)	Medium (M)	(0.3, 0.5, 0.7)
Good (G)	(0.5, 0.7, 0.9)	High (H)	(0.5, 0.7, 0.9)
Very Good (VG)	(0.7, 1.0, 1.0)	Very High (VH)	(0.7, 1.0, 1.0)

Using the linguistic variable Performance defined as in Table 1, the fuzzy decision matrix for the e-market evaluation and selection problem can be determined as

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{1}$$

Where x_{ij} represents the decision maker's assessment of the performance rating of alternative A_i with respect to criteria C_j, which is to be given by the decision maker using linguistic variables or aggregated from a lower-level decision matrix for its associated sub-criteria.

If sub-criteria C_{jk} exist for C_j, a lower-level fuzzy decision matrix can be determined in (2), where y_{jk} is the decision maker's assessment of the performance rating of alternative A_i with respect to sub-criteria C_{jk} of the criteria C_j.

$$Y_{C_j} = \begin{bmatrix} y_{11} & y_{21} & \cdots & y_{n1} \\ y_{12} & y_{22} & \cdots & y_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ y_{1p_j} & y_{2p_j} & \cdots & y_{np_j} \end{bmatrix} \tag{2}$$

The weighting vectors for the evaluation criteria C_j and sub-criteria C_{jk} can then be given in (3) and (4) by the decision maker using the linguistic variable Importance defined in Table 1.

$$W = (w_1, w_2, ..., w_j) \tag{3}$$

$$W_j = (w_{j1}, w_{j2}, ..., w_{jk}) \tag{4}$$

With the formulation of the lower-level fuzzy decision matrix for criteria C_j in (2), and the weight vector in (4) for its associated sub-criteria C_{jk}, the decision vector $(x_{1j}, x_{2j}, \ldots, x_{nj})$ across all the alternatives with respect to criteria C_j in (1) can be determined by

$$(x_{1j}, x_{2j}, ..., x_{nj}) = \frac{W_j Y_{Cj}}{\sum_{k=1}^{p_j} w_{jk}} \tag{5}$$

With the e-market selection and evaluation problem described as above, the overall objective for solving the e-market evaluation and selection problem is to rank all the alternative e-markets by giving each of them an overall performance rating with respect to all criteria and their associated sub-criteria. The process of determining the overall performance of each alternative e-market across all the selection criteria and their associated sub-criteria starts with calculating the overall weighted performance matrix of all the alternatives with respect to multiple evaluation and selection criteria by multiplying the criteria weights w_j and the alternative performance rating x_{ij}, shown as follows:

$$Z = \begin{bmatrix} w_1 x_{11} & w_2 x_{12} \cdots & w_m x_{1m} \\ w_1 x_{21} & w_2 x_{22} \cdots & w_m x_{2m} \\ \ldots & \ldots \quad \ldots & \ldots \\ w_1 x_{n1} & w_2 x_{n2} \cdots & w_m x_{nm} \end{bmatrix} \tag{6}$$

To avoid the complex and unreliable process of comparing fuzzy utilities often required in fuzzy multi-criteria analysis (Deng and Yeh, 2006), the defuzzification method determined by (7) based on geometric centre of a fuzzy number, is applied to the weighted fuzzy performance matrix in (6) (Chen and Hwang, 1992).

$$r_{ij} = \frac{\int_{S_{ij}} x \mu_{w_j x_{ij}}(x)\, dx}{\int_{S_{ij}} \mu_{w_j x_{ij}}(x)\, dx} \tag{7}$$

Where S_{ij} is the support of fuzzy number $w_j x_{ij}$ in (6). For a triangular fuzzy number *(a, b, c)*, (7) is simplified as (8)

$$r_{ij} = \frac{a+b+c}{3} \tag{8}$$

A weighted performance matrix in a crisp value format can then be obtained as

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{9}$$

To rank the alternatives based on (9), the TOPSIS method is applied. To facilitate the use of the TOPSIS method, the concept of the positive-ideal and the negative-ideal solution is used. The positive-ideal solution A^+ and the negative-ideal solution A^-, representing the best possible and the worst possible results among the alternatives respectively across all criteria, can be determined by

$$A^+ = (r_1^+, r_2^+, ..., r_m^+), \quad A^- = (r_1^-, r_2^-, ..., r_m^-) \tag{10}$$

Where

$$r_j^+ = max (r_{1j}, \ r_{2j}, ..., r_{nj}), \ r_j^- = min (r_{1j}, \ r_{2j}, ..., r_{nj}) \tag{11}$$

From (10) to (11), the distance between alternative A_i and the positive-ideal solution and between alternative A_i and the negative-ideal solution can be calculated respectively by

$$d_i^+ = \sqrt{\sum_{j=1}^{m} (r_j^+ - r_{ij})^2}; \qquad d_i^- = \sqrt{\sum_{j=1}^{m} (r_{ij} - r_j^-)^2} \tag{12}$$

A preferred alternative should have the shortest distance from the positive ideal solution and the longest distance from the negative ideal solution. As a result, an overall performance index for alternative A_i across all criteria can be determined by

$$P_i = \frac{d_i^-}{d_i^+ + d_i^-} \qquad i = 1, 2, ..., n \tag{13}$$

The larger the performance index, the more preferred the alternative.

4 An Example

This section presents an example in evaluating and selecting an e-market from four available e-markets for a SME with respect to multiple evaluation and selection criteria and associated criteria as shown in Fig.1 for demonstrating the applicability of the approach for solving the general e-market evaluation and selection problem.

To start with the e-market evaluation and selection process, the performance of each e-market with respect to the evaluation and selection sub-criteria of each criterion is determined by making the subjective assessment using the linguistic variables as presented in Table 1. Tables 2 shows the assessment results of alternative e-markets with respect to each sub-criterion.

The relative importance of the evaluation criteria and its associated sub-criteria is determined by applying the linguistic variable Importance shown as in Table 1. Table 2 shows the criteria and its associated sub-criteria weights for the e-market evaluation and selection problem.

Table 2. Assessment Results for Each Sub-Criterion

Sub-Criterion	***A_1***	***A_2***	***A_3***	***A_4***
C_{11}	VG	F	P	F
C_{12}	P	G	VG	P
C_{13}	F	VG	P	P
C_{21}	F	G	G	G
C_{22}	VG	F	F	F
C_{23}	G	G	P	VG
C_{31}	G	P	G	P
C_{32}	G	VP	VG	VG
C_{33}	P	F	G	VG
C_{41}	G	F	VG	G
C_{42}	P	G	P	F

Table 3. Criteria and Sub-Criteria Weights for E-market Evaluation and Selection

Criterion	**Linguistic Weights**	**Fuzzy Number**
C_1	H	(0.5, 0.7, 0.9)
C_{11}	VH	(0.7, 1.0, 1.0)
C_{12}	H	(0.5, 0.7, 0.9)
C_{13}	L	(0.1, 0.3, 0.5)
C_2	L	(0.1, 0.3, 0.5)
C_{21}	H	(0.5, 0.7, 0.9)
C_{22}	M	(0.3, 0.5, 0.7)
C_{23}	VH	(0.7, 1.0, 1.0)
C_3	VH	(0.7, 1.0, 1.0)
C_{31}	M	(0.3, 0.5, 0.7)
C_{32}	VH	(0.7, 1.0, 1.0)
C_{33}	H	(0.5, 0.7, 0.9)
C_4	M	(0.3, 0.5, 0.7)
C_{41}	M	(0.3, 0.5, 0.7)
C_{42}	H	(0.5, 0.7, 0.9)

Table 4. Fuzzy Decision Matrix for E-market Evaluation and Selection

	C_1	***C_2***	***C_3***	***C_4***
A_1	(0.24, 0.68, 1.38)	(0.27, 0.70, 1.49)	(0.21, 0.57, 1.32)	(0.13, 0.47, 1.35)
A_2	(0.22, 0.65, 1.55)	(0.27, 0.65, 1.47)	(0.07, 0.23, 0.85)	(0.21, 0.62, 1.63)
A_3	(0.18, 0.55, 1.27)	(0.16, 0.47, 1.20)	(0.34, 0.84, 1.63)	(0.16, 0.59, 1.44)
A_4	(0.11, 0.40, 1.08)	(0.32, 0.79, 1.53)	(0.33, 0.84, 1.50)	(0.19, 0.58, 1.58)

To construct the fuzzy performance matrix for all the alternatives with respect to multiple evaluation and selection criteria as in (1), a lower-level fuzzy performance matrix of all the alternatives with respect to sub-criteria determined from Table 2 are aggregated with respect criterion weights in Table 3 using (5). Table 4 shows the aggregated fuzzy performance matrix of alternatives with respect to e-market evaluation and selection criteria.

The overall weighted e-market performance matrix of all the alternatives with respect to e-market evaluation and selection criteria is then calculated using Table 3 and Table 4, based on (6). The fuzzy numbers in the overall weighted performance matrix are further converted into comparable crisp numbers, following (8). The results are shown in Table 5.

Table 5. Weighted Performance Matrix in Crisp Numbers

	C_1	C_2	C_3	C_4
A_1	0.61	0.33	0.68	0.41
A_2	0.65	0.32	0.38	0.50
A_3	0.54	0.25	0.90	0.45
A_4	0.44	0.35	0.86	0.48

Following the approach illustrated in (9) to (13), an overall performance index for each e-market across all criteria can be calculated shown as in Table 6.

Table 6. Performance Index and Ranking for E-market Evaluation and Selection

E-market	**Distance**		**Performance Index**	**Rank**
	$A+$	$A-$	P_i	
A_1	0.24	0.36	0.60	3
A_2	0.53	0.25	0.32	4
A_3	0.16	0.54	0.77	1
A_4	0.22	0.50	0.69	2

It is clear that alternative A_3 is the preferred choice as it has the highest performance index.

5 Conclusion

E-market has been increasingly popular in the recent decade due to the benefits that it brings into organizations, in particular SMEs. SMEs with limited technical expertise and financial support are very cautious in selecting and adopting an appropriate e-market for their electronic business. How to comprehensively evaluate and effectively select a most appropriate e-market for electronic business in SMEs considering multiple criteria becomes a critical task. The presence of subjectiveness and imprecision in the subjective decision making process further complicate the evaluation and selection process. To adequately address these problems, this paper presents a multi-criteria analysis approach for assisting SMEs in effectively evaluating and selecting the most appropriate e-market in electronic business, addressing both the nature of e-market and the specific characteristics of SMEs. An example is presented that shows the applicability of the proposed approach in solving the e-market evaluation and selection problem.

References

1. Brunn, P., Jensen, M., Skovgaard, J.: E-marketplaces: Crafting a Winning Strategy. European Management Journal 20(3), 286–298 (2002)
2. Bunker, D.J., MacGregor, R.C.: Successful Generation of Information Technology (IT) Requirements for Small/Medium Enterprises – Cases from Regional Australia. In: Proc. SMEs in a Global Economy, Wollongong, Australia (2000)
3. Buyukozkan, G.: Multi-Criteria Decision Making for E-marketplace Selection. Internet Research 14, 139–154 (2004)
4. Chen, S.J., Hwang, C.L.: Fuzzy Multiple Attribute Decision Making: Methods and Applications. Springer, New York (1992)
5. Daniel, E.M., Hoxmeier, J., White, A., Smart, A.: A Framework for the Sustainability of E-marketplaces. Business Process Management Journal 10, 277–290 (2004)
6. Delone, W.H.: Determinants of Success for Computer Usage in Small Business. MIS Quarterly 12(1), 51–61 (1988)
7. Deng, H., Yeh, C.H.: Simulation-Based Evaluation of Defuzzification-Based Approaches to Fuzzy Multiattribute Decision Making. IEEE Transactions on Systems, Man, and Cybernetics 36(5), 968–977 (2006)
8. Deng, H., Yeh, C.H., Robert, J.W.: Inter-Company Comparison Using Modified TOPSIS with Objective Weights. Computers and Operations Research 27(10), 963–973 (2000)
9. Dou, W., Chou, D.C.: A Structural Analysis of Business-to-business Digital Markets. Industrial Marketing Management 31, 165–176 (2002)
10. Grieger, M.: Electronic Marketplaces: A Literature Review and a Call for Supply Chain Management Research. European Journal of Operational Research 144, 280–294 (2003)
11. Hopkins, J.L., Kehoe, D.F.: The Theory and Development of a Relationship Matrix-Based Approach to Evaluating e-Marketplaces. Electronic Markets 16, 245–260 (2007)
12. Milliou, C., Petrakis, E.: Business-to-business Electronic Marketplaces: Joining a Public or Creating a Private. Journal of Finance and Economics 9, 99–112 (2004)
13. Ravichandran, T.: Organizational Assimilation of Complex Technologies: An Empirical Study of Component-Based Software Development. IEEE Transactions on Engineering Management 52(2), 249–268 (2005)
14. Standing, C., Lin, C.: Organizational Evaluation of the Benefits, Constraints and Satisfaction with Business-To-Business Electronic Commerce. International Journal of Electronic Commerce 11, 107–135 (2007)
15. Stockdale, R., Standing, C.: A Framework for the Selection of Electronic Marketplaces: A Content Analysis Approach. Internet Research: Electronic Networking Applications and Policy 12, 221–234 (2002)
16. Tetteh, E., Burn, J.: Global Strategies for SME-Business: Applying the Small Framework. Logistics Information Management 14, 171–180 (2001)
17. Yeh, C.H., Deng, H., Chang, Y.H.: Fuzzy multicriteria analysis for performance evaluation of bus companies. European Journal of Operational Research 126(3), 459–473 (2000)

A WebGIS and GRA Based Transportation Risk Management System for Oversea Oil Exploitation

Tijun Fan[1] and Youyi Jiang[2]

[1] School of Business, ECUST, Shanghai, China
fantijun@yahoo.com.cn
[2] School of Business, ECUST, Shanghai, China
friendship_jiang@163.com

Abstract. Rapid economic growth pushes an increasing demand for China's oil resources. With China's growing dependence on oil imports, oversea oil exploitation will influences overall layout of the country's economic development and energy strategies. Because the transportation risk evaluation involves oversea complex geopolitical and economic environment, in this paper we evaluate the risk using gray relational analysis model (RGA), and develop a transportation risk management system for oversea oil exploitation combination of spatial data mining technology and WebGIS technology to provide scientific basis for safe and efficient transportation of oversea oil.

Keywords: WebGIS; Oversea oil exploitation; Gray relational analysis; Transportation risk.

1 Introduction

The rapid economic development leads to China's growing demand for energy. China domestic oil demand in 2008 is still to maintain a rapid growth, apparent consumption of oil reached 389.65 million tons, up 6.5%; oil net imports reached 199.85 million tons, up 12.5%; oil import dependency rose further, reaching 51.3%. Oil import dependency from 1995 to 2008 is shown in Fig.1 [1, 2]. As many countries starting to speed up the layout of energy strategy; the reasonable evaluation of oversea oil exploitation has become one of the pressing research topics.

Oversea oil exploitation process involves a number of factors, but China oil imports are mainly dependent on marine transportation [3]. Maritime transport faces many uncertainties related to a number of factors in the process via foreign ports arriving at national port. So establishing a transportation risk management system can effectively support the development of rational decision-making of oil transport strategy.

Because the physical meanings of the various elements are very different in transportation risk assessment and the samples are lack of regularity we use the gray relational analysis to assess the risk of transport in this system. Gray relational analysis (GRA) is a gray system theory, one of the most basic ways. GRA is equally applicable to make a systematic analysis than mathematical statistical methods when we don't know the number and laws of the sample [4]. And the results of quantitative analysis do not appear inconsistent with the results of qualitative analysis. We can see that it is in line with the transport risk assessment of oversea oil exploitation.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 138–145, 2010.

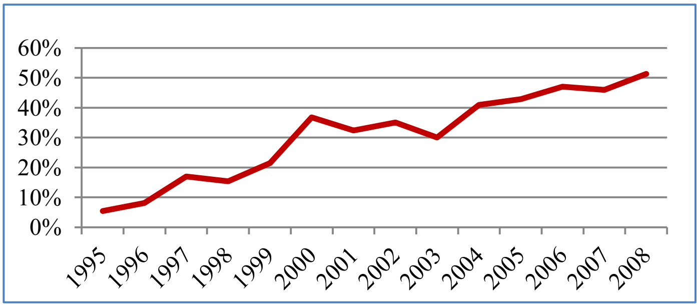


Fig. 1. China Oil import dependency from 1995 to 2008

Displaying the results of the system assessment to the users friendly must be taken into account in system development process. As an important pillar of the Earth system science Geographic Information Systems (GIS) technology has a good application in a variety of regional risk study because of its rich data analysis capabilities, particularly spatial analysis function of in regional geographic data. WebGIS is a GIS running on Internet directly, which including network, telecommunications, object-oriented, database, distributed computing, and continue to develop as the progress of these technologies [5]. WebGIS system can transport risks in a friendly and intuitive way to show to users. It is new attempting to show and analysis the risks of oil transportation with visual WebGIS query technology.

The research object of the system is the various risks involved during overseas oil transportation. In the establishment of transport risk assessment analysis model and evaluation system, based on .net framework we develop a transport risk management system for oversea oil exploitation in which using the gray relational analysis model build an evaluation model library, using a WebGIS system macro-manage and present information, and using SQL Server database with data mining technology micromanage a variety of risk indicators. We will get transport options for optimal decision-making in this system.

In this paper, how the system which includes four layers is designed will be introduced in Section 2, and in Section 3 will expound how application service layer and evaluation model layer work. At last some conclusions are given.

2 System Design

The system uses Web-based multi-layer B/S structure, namely: Web service layer, application service layer, evaluating model layer, database service layer, shown in Figure 2.

The first layer is web service layer. Web browser is used for transportation risk query, transport-related information display and transportation risk report output. As the system uses B/S structure of thin-client type, users who do not need to install any software can directly access by web browsers; Web server is the core of application functionality in the entire system which accepts and deals with requests web browsers

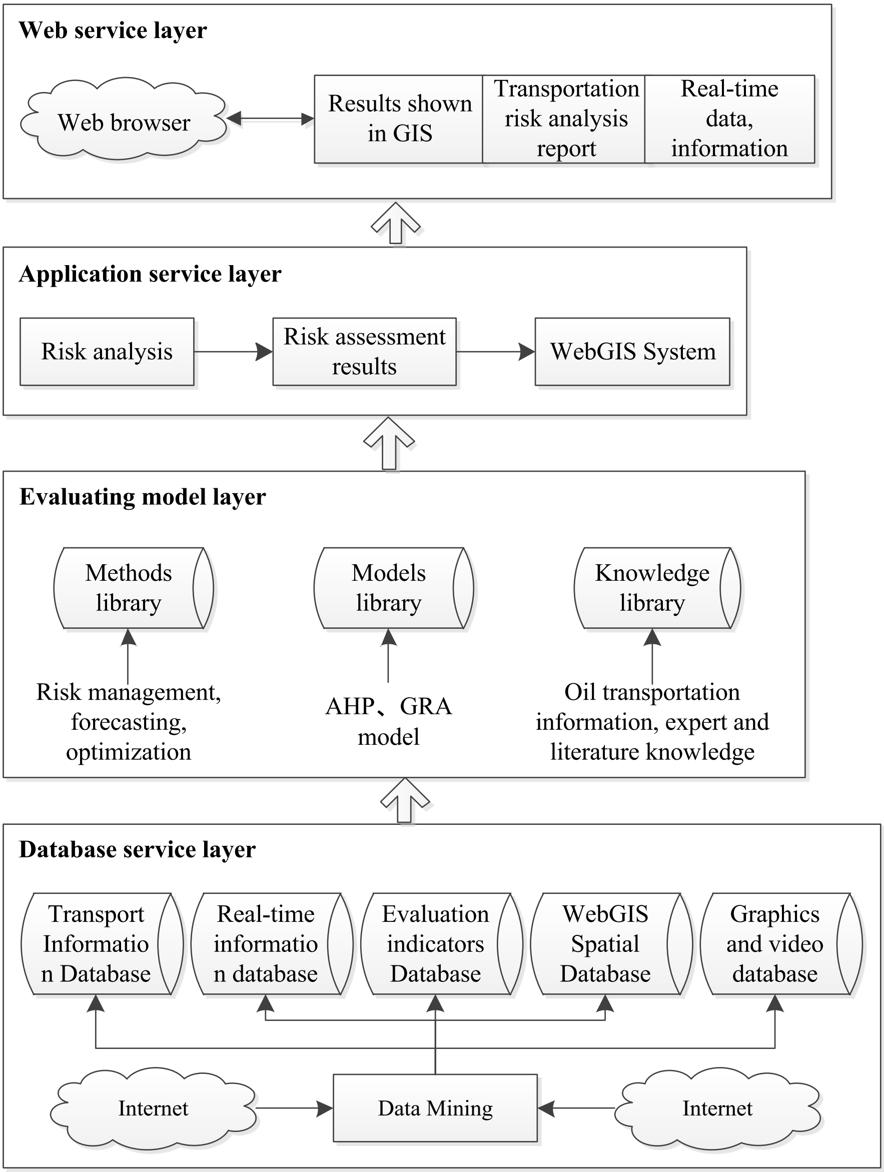

Web service layer
Web browser
Results shown in GIS
Transportation risk analysis report
Real-time data, information
Application service layer
Risk analysis
Risk assessment results
WebGIS System
Evaluating model layer
Methods library
Models library
Knowledge library
Risk management, forecasting, optimization
AHP、GRA model
Oil transportation information, expert and literature knowledge
Database service layer
Transport Informatio n Database
Real-time informatio n database
Evaluation indicators Database
WebGIS Spatial Database
Graphics and video database
Internet
Data Mining
Internet

Fig. 2. Structure of the transportation risk management system

send, and then sends the results to the web client. By JSP, ASP and other scripting languages combined with HTML, we can greatly extend the server's web features.

The second layer is the application service layer and the third layer is transportation risk evaluating model (details in part 3 Design Principles).

The fourth layer is database service layer, concentration of main data include oil transportation information data, WebGIS spatial data, transportation risk assessment indicator data, oil transportation real-time information data, graphics and video multimedia data.

3 Design Principles

3.1 Application Service Layer Design

Application server is a WebGIS server software developed using MapXtreme component, which has a large concentration of transportation risk associated with the geographic information logic and its complexity directly effects on the system's operating efficiency. WebGIS server communicates with client to deal with client's service request through web server.

Application service layer for information display and risk management of overseas oil transportation includes the results analysis of transportation risk, processing and release of transportation elements (ports, routes, pirate attacks, etc.) in the GIS spatial information platform. The application which has good expansibility is coded by C# and vb.net programming language based on the Microsoft .NET framework in the Microsoft Visual Studio 2008 development environment. Application service layer responds to client requests via web server.

In this system, WebGIS, an important function of the application service layer, achieves online query, publishing and other business processes about risk-related information on overseas oil transportation based on completing information dissemination, data sharing, exchange and cooperation on internet.

We develop WebGIS function using server-side component technology and MapInfo's MapXtreme 2008 for secondary development. As a result clients that do not need to download and install other controls and plug-ins will access WebGIS through browser which achieves interactive features with the users by JS script. Server-side developed with web development components that MapXtreme 2008 provides under Microsoft .NET Framework combined with C# language can achieve, such as data binding, selecting the layers, adding layers, zoom, spatial data query, transportation risk display, making the theme maps, custom legend and a series of complex WebGIS applications.

3.2 Evaluation Model Layer Design

Construction of the transportation risk evaluation indicator. China's crude oil imports were mainly concentrated in the Middle East and Africa, while rapid growth in South America and Asia-Pacific region, so the building of transportation risk assessment indicator requires a comprehensive, systematic and feasibility. In this paper a systematic analysis of the transport process shows that the whole transport risk includes the risks of ports, routes and pirate attacks. According to these three factors and its influence factors, oil transportation risk evaluation system is built by 3 first-level indicators and 14 second-level indicators shown in Table 1[3, 6].

Table 1. Indicator data about five representative routes[1]

First level	Second-level: evaluation indicators	Middle East routes	South America route	North Africa route	East Africa route	South-West Africa route
Port risk	Country Integrated Risk	0.537	0.078	0.217	0.217	0.217
	Throughput (T/h)	60000	1300	1000	1100	1500
	Maximum diameter of pipeline (mm)	609.6	406.4	200	203.2	406.4
	Maximum berthing tonnage (Million tons)	50	10	3.5	6.5	10
	Ability level of security	4	3	5	3	3
Route risk	Accident rate	0.326	0.459	0.343	0.418	0.315
	Route length	6000	10000	6500	6000	7000
	The number of risk nodes	4	6	5	4	5
Pirate attack risk	The number of pirate attacks in 2004	115	156	119	113	139
	The number of pirate attacks in 2005	102	117	99	124	106
	The number of pirate attacks in 2006	99	124	101	102	107
	The number of pirate attacks in 2007	78	87	80	98	112
	The number of pirate attacks in 2008	62	76	152	79	102

Transportation Risk Assessment Model Construction. According to the above evaluation system established in this paper, gray relational analysis is used for the regional transportation risk study.

Determine the weight of evaluation indicators. There are a variety of methods to determine the weights, AHP and expert scoring methods are mainly used to determine the indicator weight in the system.

Suppose there are experts respectively scoring for three first-level indicators, scores recorded as p_{ha} (h=1, 2, 3), $0<p_{ha}<1$, $\sum_{h=1}^{3} p_{ha} = 1$. Then determine the weight of each indicator by calculating the average of all the scoring value that the experts give, recorded as $A_h=\{P_1,P_2,P_3\}$, $P_h = \frac{1}{a}\sum_{a=1}^{a} p_{ha}$ (h=1,2,3).Similarly, we are still using expert scoring method and then average the scores to determine the weights of second-level indicators that influence the first- level .

At last we get the final weight of the second-level indicators, recorded as $B_j(0 <$ *B*j*<1)*,j=1nBj=1,n is the number of the second-level indicators.

Structure eigenvalue matrix of indicator. Suppose the number of the evaluation objects is m and the number of the evaluation indicators, then eigenvalue matrix of indicators can be expressed as:

[1] The port data sources from http://article.bridgat.com/guide/trans/port/port.html and other related websites; Routes and pirate raids data is gotten by collating according to data from IMB and IMO.

$$x = (x_{ij})_{m\times n}, x_{ij}(i = 1,2,3\ldots m; j = 1,2,3\ldots n) \tag{1}$$

The dimensionless of eigenvalue. As the physical meanings of indicators are different, the dimension of the data is not the same. As a result, It is not easy to compare, or difficult to obtain a correct conclusion in the comparison.

Therefore, when carrying out the GRA, the data collected must be non-dimensional treated.

For positive indicators, namely, the greater the index value, the greater the risk, its processing method is:

$$[X_{ij} - \mathrm{MIN}(X_{ij})]/[\mathrm{MAX}(X_{ij}) - \mathrm{MIN}(X_{ij})] \tag{2}$$

For the negative indicators, namely, the smaller the index value, the greater the risk, its processing method is:

$$[\mathrm{MAX}(X_{ij}) - X_{ij}]/[\mathrm{MAX}(X_{ij}) - \mathrm{MIN}(X_{ij})] \tag{3}$$

Comparative sequence and reference sequence. According to the above we have *m* objects in which there are n indicators, and then the comparative sequence is:

$$X_i = \{X_{ij} | i = 1,2,\ldots,m; j = 1,2,\ldots,n\} \tag{4}$$

Make all the optimal value (the lowest risk) of evaluation indicators as the reference sequence, denoted by X_0:

$$X_0\{X_{0j} | j = 1,2,3,\ldots,n\}, X_0 = \{X_{01}, X_{02}, X_{03} \ldots X_{0n}\} \tag{5}$$

Calculation of correlation coefficient.

$$\gamma(X_0(j), X_i(j)) = \frac{\min_i \min_j |X_{0j} - X_{ij}| + \zeta \max_i \max_j |X_{0j} - X_{ij}|}{|X_{0j} - X_{ij}| + \zeta \max_i \max_j |X_{0j} - X_{ij}|} \tag{6}$$

Calculation of correlation. The correlation between the *i* evaluation transportation route (X_i) and reference sequence(X_0) as follows:

$$\gamma(X_0\ , X_i) = \frac{1}{m}\sum_{j=1}^{m} B_j \gamma(X_0(j), X_i(j)) \tag{7}$$

Evaluation and analysis. Sort the various transport routes According to the size of the correlation. The reference sequence is constituted by the lowest risk values of the indicators. Then the larger correlation, the lower risk; the smaller correlation, the higher risk. Scored by experts the final weight matrix of all the second-level assessment indicators as follows:

$$B = \begin{Bmatrix} 0.0375, 0.05, 0.05, 0.05, 0.0625, 0.105, 0.09, 0.105, \\ 0.0225, 0.045, 0.0765, 0.126, 0.18 \end{Bmatrix}$$

We select five representative routes from a number of routes to have a specific analysis described in Table 1.

We get indicators eigenvalue matrix as follows:

$$\begin{bmatrix} 0.537 & 60000 & 609.6 & 50 & 4 & 0.326 & 6000 & 4 & 115 & 102 & 99 & 78 & 62 \\ 0.078 & 1300 & 406.4 & 10 & 3 & 0.459 & 10000 & 6 & 156 & 117 & 124 & 87 & 76 \\ 0.217 & 1000 & 200 & 3.5 & 5 & 0.343 & 6500 & 5 & 119 & 99 & 101 & 80 & 152 \\ 0.217 & 1100 & 203.2 & 6.5 & 3 & 0.418 & 6000 & 4 & 113 & 124 & 102 & 98 & 79 \\ 0.217 & 1500 & 406.4 & 10 & 3 & 0.315 & 7000 & 5 & 139 & 106 & 107 & 112 & 102 \end{bmatrix}$$

Second, we have the adjusted eigenvalue matrix after dimensionless as follows:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0.5 & 0.076 & 0 & 0 & 0.047 & 0.12 & 0 & 0 & 0 \\ 0 & 0.995 & 0.496 & 0.860 & 1 & 1 & 1 & 1 & 1 & 0.72 & 1 & 0.265 & 0.156 \\ 0.303 & 1 & 1 & 1 & 0 & 0.194 & 0.125 & 0.5 & 0.140 & 0 & 0.08 & 0.059 & 1 \\ 0.303 & 0.998 & 0.992 & 0.935 & 1 & 0.715 & 0 & 0 & 0 & 1 & 0.12 & 0.588 & 0.189 \\ 0.303 & 0.992 & 0.496 & 0.860 & 1 & 0 & 0.250 & 0.5 & 0.605 & 0.28 & 0.32 & 1 & 0.444 \end{bmatrix}$$

Third, according to correlation coefficient formula we obtain correlation coefficient matrix:

$$\begin{bmatrix} 0.333 & 1 & 1 & 1 & 0.5 & 0.868 & 1 & 1 & 0.914 & 0.806 & 1 & 1 & 1 \\ 1 & 0.334 & 0.502 & 0.368 & 0.333 & 0.333 & 0.333 & 0.333 & 0.333 & 0.410 & 0.333 & 0.654 & 0.762 \\ 0.623 & 0.333 & 0.333 & 0.333 & 1 & 0.720 & 0.8 & 0.5 & 0.781 & 1 & 0.862 & 0.894 & 0.333 \\ 0.623 & 0.334 & 0.335 & 0.348 & 0.333 & 0.411 & 1 & 1 & 1 & 0.333 & 0.806 & 0.460 & 0.726 \\ 0.623 & 0.335 & 0.502 & 0.368 & 0.333 & 1 & 0.667 & 0.5 & 0.452 & 0.641 & 0.610 & 0.333 & 0.530 \end{bmatrix}$$

Last, we get the correlation between the evaluation transportation routes and reference sequence shown in Table 2.

Table 2. Correlation of Five Routes

Transportation risk	Correlation	Ranking
Middle East route	0.9193	1
South America route	0.4895	5
North Africa route	0.6372	2
East African route	0.6210	3
South-West Africa route	0.5449	4

Therefore, in oil transportation process risk ranking of the five exam routes from high to low are: South America route, South-West Africa route, East African route, North Africa route and Middle East route.

4 Conclusions

This paper describes a method how to develop a B/S mode based oversea oil transportation risk decision support system. In the system development, the successful use of the COM components, JavaScript, WebGIS, database management and a series of techniques makes the entire system have a high portability and maintainability. In the actual system application process, we find oil transportation risk assessment has a good guide decision-making for users with high accuracy and real-time. At present, this technology has been successfully applied in risk management system for oversea mineral resources exploitation & utilization which received praise from many experts.

The main tasks of the next study are two aspects. First, we should further improve the risk assessment system of oil transportation, such as the impact of risk about the choice of oil carriers, so that the system assessment and decision- making are more scientific and rigorous; Second, we should combine data mining and data warehouse technology to further improve the intelligent spatial analysis and query capabilities of WebGIS system, so that WebGIS will play a more important role in visual query of oil transportation risk at different time and space.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (70871038), Innovation Program of Shanghai Municipal Education Commission (09ZS65), Shanghai Talent Development Fund (2009018) and the Fundamental Research Funds for the Central Universities.

References

1. Gao, H., Ding, H., Li, L.: Current Situation of Petroleum Consumption in China and Its Strategic Thinking. China Safety Science Journal 14, 29–33 (2004)
2. Minxuan, C.: China's Energy Development Report. Social Sciences Academic Press, Beijing (2008)
3. Wu, G., Wei, Y.: China's oil import risk analysis of maritime transport. Energy of China 31, 9–12 (2009)
4. Deng, J.: Gray Theory Basic. Huazhong University of Science and Technology Press, Wuhan (2002)
5. Theseira, M.: Using Internet GIS technology for sharing health and health related data for the West Midlands Region. Health Place 8, 37–46 (2002)
6. Sun, J., Jia, D.: China's oil security system and evaluation of maritime transport (I). Shipping Management 27, 22–25 (2005)

Prediction of Pine Wilt Disease in Jiangsu Province Based on Web Dataset and GIS

Mingyang Li[1,*], Milan Liu[1], Min Liu[1], and Yunwei Ju[2]

[1] Department of Forest Management, Nanjing Forestry University, Nanjing 210037, China
{Li,Liu,Liu}lmy196727@126.com
[2] Department of Forest Protection, Nanjing Forestry University, Nanjing 210037, China
jyw6808@yahoo.com.cn

Abstract. 80 pine wilt disease occurrence points with geographical coordinates in 2007 and 31 environmental variables from open web datasets were gathered as the main source of information. Four modeling methods of Classification and Regression Trees (CART), Genetic Algorithm for Rule-set prediction (GARP), maximum entropy method (Maxent), and Logistic Regression (LR) were introduced to generate potential geographic distribution maps of pine wood nematode in Jiangsu province, China. Then we calculated three statistical criteria of area under the Receiver Operating Characteristic Curve (AUC), Pearson correlation coefficient (COR) and Kappa to evaluate the performance of the models. The results showed that: CART outperformed other three models; slope, precipitation, seasonal variations (bio15), mean temperature of driest quarter (bio9), north-south aspect (northness), maximum temperature of warmest month (bio5) were the six enforcing environmental factors; future occurrence area of pine wilt disease will be 47.27% of total pine forest, tripling present infected area of the pest.

Keywords: pine wilt disease; prediction; web dataset; GIS.

1 Introduction

Pine wood nematode *Bursaphelenchus xylophilus*, which belongs to Nemathelminthes division, is a destructive alien pest to pine forest. It is about one thousandth centimeter long and not easy to be seen by naked eyes of human beings. As early as in 30th of 20 century, it was first reported as a new species by Steiner and Buhren [1], but was not regarded as the major cause of pine wilt disease until 60th of the last century [2]. Pine wood Nematode is not widely distributed in the world at present time, mainly concentrated in US, Canada and Mexico of America, China, South Korea and Japan of Northeast Asia, Portugal of Europe. However, it does not bring destructive damages to forest in America, while causing death of millions of pine trees in Asian countries. In 1982, pine wood nematode was first discovered in Dr. Sun Yat-sen Mausoleum of Nanjing, Jiangsu Province, China. By the end of 2002, the alien pest has spread to over 12 provinces of China, including Anhui, Zhejiang,

[*] Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 146–153, 2010.

Guangdong, Jiangxi, Hubei, etc. Since 1982, over 5 million pine trees in Jiangsu province have been dead, which causes timber losses of 250 thousand cubic meters and economical losses of 200 million *yuan* in RMB [3].The widespread of the exotic species has severely impacted the social and economical sustainable development of the infected areas.

As one of the most severe forest pests, simple, practical and economical prediction method of pine wood nematode has not been found yet. "Giving priority to prevention and comprehensive treatment", is the guideline of China's forest protection and precious experience learned from many years practice. Therefore, it is imperative to predict potential habitat of pine wood nematode and to understand ecological determinants of spatial patterns of the pest.

Since enforcing variables impacting potential habitat of species are diversified, it is very difficult to model habitat for invasive species by means of traditional field survey methods. With the advancement of applied mathematics and computer science, many ecologists begin to apply statistical method and Geographical Information System (GIS) technology to model spatial distribution of invasive species, and have developed several ecological niche programs, such as CLIMEX, BIOCLIM, GARP, and so on. Kelly et al. precisely modeled spatial distribution of Sudden Oak Death in Middle West of US based on Support Vector Machine (SVM) [4]. Bendor et al. applied spatial explicit model to predict spatial distribution of Emerald Ash Borer [5]. Howell et al. researched spatial distribution of White Pine Blister Rust (WPBR) in Colorado, USA, based on model of CART [6].

In China, many research papers on pine wood nematode have been published. However, these documents were mainly focused on such aspects as the mechanism of the disease, interaction between nematode and vector beetle, ecological impacts and prevention measures. Just a few scientists have done some single modeling research on the potential habitat of the exotic species, lacking the comparison between multiple models and detailed analysis of contributing environmental factors. Furthermore, prediction of occurrence area of the pest based on potential habitat has not been done by Chinese forest protection experts.

There were three objectives in the paper: (1) to evaluate the performance of four ecological niche models, namely CART,GARP, Maxent and LR; (2)to analyses the enforcing factors impacting the spatial distribution of pine wood nematode; and (3) to predict the infected area of pine wilt disease in Jiangsu Province by county or district.

2 Materials and Methods

2.1 Study Area

Jiangsu province is located in east coast of China (116°22′～ 121°55′E, 30°46′～ 35°07′N), with a span of 460 km from north to south and 320 km from east to west. It is bounded by Huang Sea in the east, Shandong Province in the north, Anhui Province in the west, Zhejiang Province and Shanghai City in the southeast. It covers an area of approximately 102,600 km^2, in which plain, water, mountain and hill comprise 69%, 17%, 15%, respectively. Jiangsu is one of the most populous provinces in China, with its 2006 population at approximately 75.49 million people. The province lies in the transition belt between subtropical zone and warm temperature zone, with moderate climate and clearly seasonal change, which can provide excellent natural conditions for vegetation to grow. The regional climax forest type is deciduous broadleaf forest, with only a few evergreen species in the southern part. Man made plantation comprises over 96% of total forest area, in which there are only a handful tree species, mainly pines, Chinese fir, bamboo, cypress and poplar. Sparely distributed forest, rapidly development of foreign trade, huge population, unreasonable forest structure, combined together make forest ecosystem become very susceptible to the invasion of alien forest pests.

2.2 Species Occurrence Data

Occurrence data of pine wood nematode are sparse and there are no complete online digitalized presence and absence data with geographical coordinates at provincial scale. In addition, even when absence data in some areas are available, they may be of questionable value in many situations. We collected 80 occurrence place data of pine wood nematode in 2007 from local county level forest bureaus of Xinpu, Yuntai and Jiangning, and then assigned geographical coordinates to these data from gazetteers in Jiangsu GIS dataset. Finally, we acquired 80 occurrence data with coordinates.

2.3 Environmental Data

In the paper, we assumed that habitat variables of pine wood nematode were related to climate, terrain, tree cover and soil. Exacting terrain and forest cover data from topographic maps and atlases over large area is time-consuming and expensive. Instead we downloaded free online digitalized data from open web datasets and managed them using ArcGis version 9.2 software. The 31 environmental variables used were based on documented species–habitat associations (Table 1).

Table 1. Environmental Variables of Pine Wood Nematode

Category	Variables	Data source
Climate	bio1~bio19	WORDCLIM
Terrain	slope, northness: cos(aspect), eastness:sin(aspect), CTI: compound topographic index, elev: elevation, flowdir: flow direction, flowac: flow accumulation	EROS
Land cover	tree_cover: tree cover NPP: Net Primary Production in 2000	GLCF
Soil	soil_type: soil type, soil_ph: soil pH, soil_moiture: soil moisture	Asian HYDRO 1K dataset

Terrain data of slope and aspect were downloaded from data center of Earth Resource Observation and Science (EROS), USGS (http://edc.usgs.gov/products/ elevation /gtopo30/ hydro/asia.html) at a resolution of 1 km per pixel. Climate data of 19 variables and elevation were acquired from dataset of WorldClim version

1.4 (1950-2000) (http://www. worldclim.org/) at a resolution of 30 seconds per pixel (0.93 x 0.93 = 0.86 km^2 at the equator). Tree cover data of Nov.2000 to Nov.2001, net primary production (NPP) of 01-01-2000 to 12-31-2000 were taken from the Global Land Cover Facility (GLCF) (http://www.landcover. org, accessed in 2007) at a resolution of 500 m per pixel and 90 m per pixel respectively in the projection of Longitude/Latitude WGS84. Soil data of soil type, soil moisture and soil pH were downloaded from Asian HYDRO 1K dataset at a resolution of 1 km per pixel in projection of Lambert Azimuthal Equal Area. All these data were clipped with the mask of Jiangsu Province, re-sampled to a 30 second pixel size, re-projected to the same projection of Longitude/ Latitude WGS84 on platform of ArcGis 9.2.

2.4 Data Preparation

There are complex correlations among 31 environmental variables of pine wood nematode, which makes it more difficult for people to identify the contributing environmental factors. On the other hand, correlations among different variables could not meet the necessary preconditions of LR model of normal distribution and independence of variables. So, a Pearson correlation analysis of 31 environmental variables of 80 occurrence points was done on SYSTAT. Afterwards, Maxent model was run to provide contribution percent for 31 environmental factors. Among correlated variables, those with bigger contribution percent were remained to build four prediction models. Altogether, we picked out six ecological factors which added up 91.3 % of cumulative contribution rate as following: slope （20.2%） , precipitation seasonality （bio15, 12.9%） , compound topographic index (CTI, 8.2%), mean temperature of driest quarter (bio9,7.8%), north-south aspect (northness,2.8%), maximum temperature of warmest month (bio5,1.6%).

Occurrence points are enough to run the two models of Maxent and Garp. However, another data of independent absence points are necessary to build LR model and CART model. Besides, absence points are also used in the procedure of model validation. Generally, background points beyond presence points are assumed as absence points. On the platform of ArcGis, 2,000 random sampling points were generated in the territory of Jiangsu Province. To prevent occurrence of aggregation, a minimum distance of 0.1° （GCS_WGS_1984 projection） between two points was set. Intersect Point Tool of HawthTools was applied to extract potential habitat probability of 2,000 random sampling points on the Maxent distribution map. According to the principles of statistics, if a point is with a very small occurrence probability ($p<0.05$), we can assume this point is an absence. Thus we got 1,789 absence points of pine wood nematode. A uniform database was created by merging two shape files of presence points and absence points. Then we assumed the code of presence point as ‘1.0’, absence points ‘0’. Seventy percent points (1,308) from the uniform database were randomly selected to generate models, while the remaining 30 percent points (561) to test prediction accuracy of the four models.

The ecological model of best overall performance was selected to make prediction of pine wilt disease in Jiangsu Province, China. Based on the theory of statistical probability, the pixel value was set to be ‘1’ when occurrence probability was more than 0.95; else ‘0’. According to the technical manual of Chinese Forest Inventory and Design, when tree cover (canopy density) in a stand is equal to or above 0.2, the stand type can be classified into forest. After raster layer of tree_cover of Jiangsu Province was loaded into ArcGis environment, pixel value was set to be ‘1’ when tree cover>0.2; else ‘0’. Supported by Raster Calculator Tool, we did logic operation AND between two raster layers of pine wood nematode occurrence and tree_cover, then raster layer of pine wilt disease was generated. Zonal statistic calculation was done by county (district) and the result of predicted area of pine wilt disease in Jiangsu province was figured out.

3 Results

3.1 Models Development

Six environmental variables were chosen to build four habitat predication models. As for CART, among six environmental variables, only three factors of slope, bio9, bio15 operated in the model. With a PRE value of 0.905, 90.5% of environmental variance could be interpreted. By using conditional judgment function of con( ) in the Arcmap tool of Raster Calculator, we made a habitat distribution map (Figure 1-a) after following parameter inversion equation had been built:

$$\mathrm{CART}=\mathrm{con}(\mathrm{slope}<15.6,\ 0.011,\ \mathrm{con}\ (\mathrm{bio9}>6.7,0,\mathrm{con}\ (\mathrm{bio15}>55,\ 0.258,0.971))) \tag{1}$$

For Maxent model, the purpose of the first execution was to identify the six contributing environmental variables and to help generate absence points for the exotic species. When Maxent model was run for the second time, only six variables were input into model. The training accuracy of the Maxent model was 0.914 (Figure 1-b).

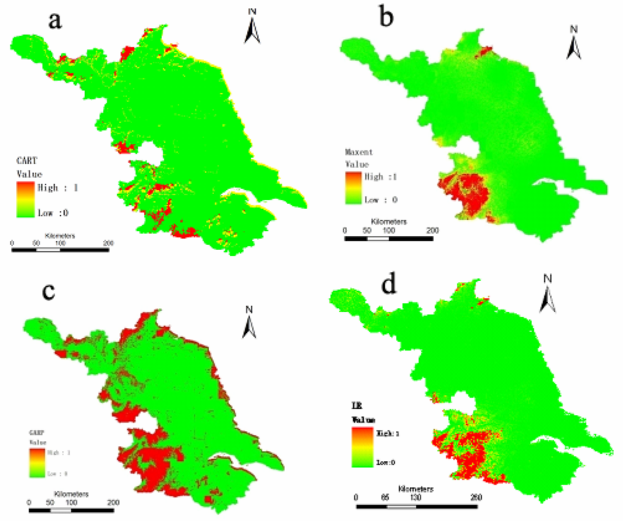


Fig. 1. Potential Habitat Maps for Pine Wood Nematode

Since Desk Garp version 1.1.6 is not stable and the result of each task is different [7], we built Garp model with default values for Desktop Garp (20 tasks, 0.01 convergence limit, 1,000 maximum iterations). We selected the task with the highest training accuracy (0.872) from the best subset as the prediction result (Figure 1-c).

As for LR model, variables of bio5, bio15 and slope were natural logarithm, sine transformed respectively, to meet the normal distribution condition of the model. The prediction model of LR was generated as following:

$$y=-31.133-0.004\times CTI+2.338\times Sin(slope)+30.882\times Log(bio5)-24.835\times Log(bio15)+0.264\times northness-0.050\times bio9 \quad (2)$$

In which, Log and Sin stand for data transformation of common logarithm and sine, respectively. With a Naglekerke's R^2 value of 0.89, LR model could interpret 89% of environmental variance, which indicates high simulation performance of the model. By using the tool of Raster Calculator on the platform of ArcGis, we built a parameter inversion equation of LR= Exp(y) / [Exp(y) +1] and made a potential habitat distribution map for pine wood nematode (Figure 1-d).

3.2 Models Validation

Three statistics of AUC, Kappa and COR were applied to asses the prediction performance of four models (Table 2). It can be seen from Table 2, AUC values of the four models were all above 0.8, showing the excellent prediction performance of the selected models. For three statistics, there were no distinct difference between different models and the average of three indexes for each model was approximate. Among the four models, CART ranks the first in terms of P-Kappa, COR and average. Therefore, the result of CART model was applied to make prediction of pine wilt disease in Jiangsu Province.

Table 1. Evaluation Statistical Criteria of Prediction Models

Criteria	CART	Maxent	GARP	LR
AUC	0.921	0.943	0.890	0.908
P-Kappa	0.849	0.838	0.807	0.818
COR	0.846	0.805	0.729	0.809
Average	0.872	0.862	0.809	0.845

3.3 Enforcing Environmental Variables

As shown in equation (1), enforcing environmental factors of pine wood nematode were ranked according to the sequence from high to low as following: slope, mean temperature of driest quarter (bio9), precipitation seasonality (bio15).It can be seen from CART model that, pine wood nematode likes to inhabit in higher hills and mountain area (slope> 15.6 °) with lower mean temperature of driest quarter (bio9 <6.7℃) and small seasonal variations in precipitation (bio15 <55%).

It can be seen from equation (2) that, potential habitat probability of pine wood nematode was positively correlated with slope, bio5 and northness, negatively correlated with CTI, bio9 and bio15. CTI is a steady state wetness index, which is a function of both the slope and the upstream contributing area per unit width orthigonal to the flow direction. CTI is highly correlated with several soil attributes such as horizon depth (r=0.55), silt percentage (r=0.61), organic matter content (r=0.57), and phosphorus (r=0.53) [8]. In Jiangsu Province, the driest season is summer, while the warmest season is spring. LR model shows that in the selection of potential habitat, pine wood nematode prefers the warm and humid spring, and can not tolerate the heat in hot and dry summer.

From the above analysis, we can conclude that pine wood nematode likes to live in the warm and subtropical zone with high annual temperature and small precipitation seasonality. As for the choice of sites, the alien pest tends to inhabit in sunny slope with barren soil condition.

3.4 Prediction of Pine Wilts Disease

In Jiangsu Province, pine trees are the dominant afforestation species in hilly and montane areas, especially in scenic regions. The total area of pine forest in Jiangsu is 93,333 ha, in which monsoon pine *Pinus massoniana* and black pine *Pinus thunbergii* are 46,667 ha and 38,000 ha respectively. The pine forest is mainly distributed in montane area of Nanjing and Liyang, coastal mountainous region of Yuntai, and ring belt of Tai Lake.

Predicted occurrence area of pine wilt disease in the future is 44,119.88 ha, which comprises 47.27% of total pine forest (93,333 ha), tripling present infected area of the pest (13,333 ha).Ranked in accordance with the order of descending probability, Yixing, Liyang and Jurong will be the counties which are most susceptible to the pest, while Yixing, Liyang and Nanjing urban district will be the regions with the largest area of infected pine forest. Combined with spatial distribution of pine trees and economic developing level, we can make a projection that, the area of pine wilt disease is proportional to the area of pine forest, and pine wilt disease is most apt to break out in the counties with large area of pine forest and booming economic.

4 Conclusions

At present, occurrence data with geographical coordinates of alien forest pest in China are not completed, and much ecological and geological information of invasive species can not be shared among different research institutions. Under this situation, modeling potential habitat for alien species based on limited occurrence data, open web datasets and GIS can provide practical and economical approach for prevention research of biological invasion.

There are three kinds of enforcing environmental factors which affect spatial distribution of pine wood nematode: climate, topography and soil. Among them, slope, CTI and aspect are the environmental factors impacting the distribution of host trees of monsoon pine and black pine, while mean temperature of driest quarter, precipitation seasonality and maximum temperature of warmest month are the major ecological factors of vector beetle of *Monochamus alternatus*. Therefore, both the environmental factors of host tree and vector insect should be considered when modeling potential habitat for pine wood nematode.

Graham regarded that changes in habitat use are more closely related to the extremes of temperature and rainfall than to the averages based on the research on the fossil record [9]. That means the differences between the wettest and driest season, and the hottest and coldest temperatures may ultimately determine the capability of many plants and animals to occupy any given area. Our research confirmed Graham's conclusion. It should be paid attention that the results of our paper are based on the specific natural environmental conditions of study area. Whether the research conclusions are applicable to other areas, still needs to be treated with caution.

It is predicted by CART model that future occurrence area of pine wilt disease would be 47.27% of total pine forest, which is equal to three times of present infected area of the pest. The prediction shows that it will take a long time before the alien forest disease is eliminated completely. Among the six enforcing ecological factors, the climatic variables are difficult to be modified by mankind in the near future. However, improving soil condition and increasing management intensity of pine forest should become the major components of the strategy for controlling pine wilt disease.

Acknowledgment. This research was funded by Key Laboratory Open Foundation of State Education Ministry (No. GJSYS-100220402) and 948 Project of State Forestry Administration, China (No.2008-04-56).We thank personnel in Forest Bureaus of Liangyungang City and Nanjing City for their enthusiasm and support, and Guangcai Xu for his field and lab assistance.

References

1. Steiner, G., Buhrer, E.M.: *Aphelenchoides xylophilus*, a Nematode Associated with Blue-stain and Other Fungi in Timber. J. Agric. Res. 48, 949–951 (1934)
2. Yang, B.J., He, C.Y., Wang, C.F.: General Occurrence Situation of Pine Wilt Disease Abroad. Forest Pest and Disease 5, 40–42 (1999)
3. Xie, C.X., Zheng, H.Y., Zhang, P.: Occurrence Causes and Management Measures of Pine Wilt Disease in Jiangsu Province. Journal of Jiangsu Forestry Science and Technology 29, 41–43 (2002)
4. Kelly, M., Meentemeyer, P.R.: Landscape Dynamics of the Spread of Sudden Oak Death. Photogramm Eng. Rem. Sens. 68, 1001–1009 (2002)
5. BenDor, T.K., Metcalf, S.S., Fontenot, L.E., Sangunettc, B., Hannonb, B.: Modeling the Spread of the Emerald Ash Borer. Ecological modeling 197, 221–236 (2006)
6. Howell, B., Burns, K., Kearns, H., Witcosky, J.J., Cross, F.J.: Biological Evaluation of a Model for Predicting Presence of White Pine Blister Rust in Colorado Based on Climatic Variables and Susceptible White Pine Species Distribution. Biol. Eval. 15 R2-06-04 (2006)
7. Stockwell, D., Peters, D.: The GARP Modeling System: Problems and Solutions to Automated Spatial Prediction. Int. J. Geographical Information Science 13, 143–158 (1999)
8. Moore, I.D., Gessler, P.E., Nielsen, G.A., Gallant, J.C.: Terrain Attributes: Estimation Methods and Scale Effects. In: Jakeman, A.J., Beck, M.B., McAleer, M.J. (eds.) Modeling Change in Environmental Systems, pp. 189–214. Wiley, London (1993)
9. Graham, A.: The Current Status of the Legume Fossil Record in the Caribbean Region. In: Herendeen, P.S., Dilcher, D.L. (eds.) Advances in Legume Systematics. Part 4. The fossil record, Royal Botanical Gardens, Kew, London, pp. 161–167 (1992)

An Approach for Integrating Geospatial Processing Services into Three-Dimensional GIS

Yingjie Hu, Jianping Wu, Haidong Zhong, Zhenhua Lv, and Bailang Yu[*]

Key Laboratory of Geographic Information Science, Ministry of Education
East China Normal University, Shanghai 200062, China
yjhu.geo@gmail.com, jpwu@geo.ecnu.edu.cn, zhd_1981@163.com,
zhenhualv@gmail.com, blyu@geo.ecnu.edu.cn

Abstract. Three-dimensional (3D) GIS is gaining more and more acceptance among both scientists and the general public. Though powerful in data visualization, 3D GIS is comparatively weak in geospatial analysis. In order to enhance 3D GIS's analysis capability, we present an approach for integrating geospatial processing services into 3D GIS. The architecture of our approach contains four layers which are the Presentation layer, the Application layer, the Service layer and the Data layer. Two different workflows are designed for this architecture to deal with geospatial tasks of different complexities. We also implement this approach in a 3D GIS project named Digital Chongming Island (DCI), Shanghai, China. By successfully integrating a variety of geospatial processing services into DCI, we have demonstrated that our approach is feasible and effective.

Keywords: 3D GIS; geospatial processing services; service and application integration.

1 Introduction

Three-dimensional (3D) GIS is gaining more and more acceptance among both scientists and the general public [1]. By visualizing spatial data in a manner that people naturally comprehend, 3D GIS makes full use of human's highly developed recognition skills, and lowers the barrier to entry for many users [2][3]. 3D GIS also represents some scientific findings in a more attractive way, thereby helping draw more attentions to these discoveries [4].

Despite the numerous conveniences, 3D GIS is still in its infancy [1]. Popular 3D GIS products, such as Google Earth, World Wind and Skyline Globe, are focusing on displaying virtual environment vividly and smoothly, and have offered few geospatial analysis functions [5][6][7]. Compared with conventional 2D GIS software such as ESRI's ArcGIS Desktop, current 3D GIS products are almost incapable to accomplish the complex tasks which require advanced geospatial analysis functions. As a result, users still have to recourse the tools of 2D GIS to complete their tasks rather than get the whole work done directly on a 3D GIS platform, and their working efficiency is consequently decreased.

[*] Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 154–161, 2010.

The emergence of web service has provided a possible way to enhance 3D GIS's geospatial analysis capability. Web service is a standard-based computing unit which is published on the Internet and helps share valuable data and programs among a large number of users [8][9]. Geospatial processing services encapsulate spatial analysis tools, and make such powerful functions accessible to a wide range of client applications including many 3D GIS platforms. By integrating geospatial processing services into 3D GIS, we can easily extend 3D GIS's analysis functions.

This paper addresses the issue of combining geospatial processing services and 3D GIS. The goal of our research is to design an architecture which enables users to perform advanced geospatial analysis in 3D environment. Our work not only focuses on integrating single service into 3D GIS, but also moves on to service chaining, the process of combining several elementary services into a service chain that can execute complex functions [10]. We also apply our approach to a 3D GIS project named Digital Chongming Island (DCI), Shanghai, China, and demonstrated that the method can effectively integrate geospatial processing services (including simple services and chained services) into 3D GIS.

The rest of the paper is organized as follows. Section 2 outlines some relevant concepts and reviews related work. Section 3 presents the detailed description of the architecture of our approach and elaborates on the key components. The implementation and application of the approach in DCI are described in section 4. Finally, we summarize this research and draw some conclusions in section 5.

2 Related Work

2.1 Geospatial Services

Geospatial services are specialized web services, and have inherited all properties of general services. What make geospatial services different from other services are the inherent spatial characteristics of the data which geospatial services have to deal with [11]. Geospatial services can be grouped into three categories: data services, processing services and registry services [10]. Since registry services are responsible for searching, maintaining and accessing other services, we only discuss the other two types of services which directly operate on geospatial data.

Data services make it possible to share geospatial data, which are often huge, complex and heterogeneous, on the Internet [12]. Typically, data services have standardized interfaces that allow users to access a customized portion of a specific dataset stored in a particular repository. Web Map Service (WMS) and Web Feature Service (WFS) are the most widely used data services which are proposed by Open Geospatial Consortium (OGC). While WMS uses maps, which are produced in a pictorial format such as PNG, GIF or JPEG, to portray geographic information, WFS allows client to retrieve geospatial data encoded in Geographic Markup language (GML).

Processing services encapsulate geospatial algorithms and provide access to sets of operations through standardized interfaces [13]. They receive requests from users, execute geospatial processing functions (e.g. projection conversion), and send the results back to users. The datasets manipulated by processing services can be either user's local data or those deposited on remote servers. Users can invoke a particular geospatial processing service without understanding its internal mechanism. Thus, processing services effectively increase the reusability of geospatial analysis functions, and made such functions accessible to a larger number of users.

2.2 Integrating Geospatial Services into Applications

In order to better share geographic data and tools, many researches have been conducted on integrating geospatial services into client applications [14][15][16]. Service-Oriented Architecture (SOA) has been adopted in most of these researches. In SOA, geospatial services, as the basic units, are collected by developers to build larger and more complex applications. Paul and Ghosh (2008), focusing on geospatial data services, present a service-oriented approach for integrating heterogeneous spatial data sources [15]. Carlos and Laura (2010) move on to geospatial processing services and have developed a service-oriented application allowing hydrologists to access both data and processing services [16]. However, all of these researches are centered on integrating geospatial services into 2D GIS applications rather than 3D GIS platforms.

The combination of 3D GIS and geospatial services has also gained attention among scientists and researchers. Dunne and Sutton (2006) have proposed ways to integrate marine data, which are published as WMS, into NASA's World Wind [17]. Craglia and Goodchild (2008) pointed out that virtual globe may become the next platform for service integration and data visualization [18]. Not only researchers and scientists, several 3D GIS software companies also improve their virtual globes to support standard geospatial services. For example, both Google Earth and Skyline Globe have provided additional modules to integrate WMS and WFS into their 3D environments [5][6]. However, most of these researches focus on integrating data services rather than processing services into 3D GIS.

3 Architecture

We present a layered architecture (Figure 1) to integrate geospatial processing services into 3D GIS. This architecture contains four layers: the Presentation layer, the Application layer, the Service layer and the Data layer.

3.1 The Presentation Layer

As Figure 1 illustrates, the Presentation layer contains two modules: the User Interface and the 3D Viewer. The User Interface is responsible for interacting with end user. It captures the parameters input by user, and utilizes its AJAX Engine to send these parameters to the Task Manager component in the Application layer. The User Interface also receives the results of the geospatial processing services from the Task Manager, and redirects such results to the 3D Viewer. The 3D Viewer can be an ActiveX control, like Google Earth Plug-in or Skyline Globe 3D window, which is employed to retrieve the geospatial data (e.g. DEM, images, and 3D models) from the Data layer, and to render such data as 3D scene. The 3D Viewer also visualizes the results of the geospatial processing services, and integrates the data services (e.g. WMS and WFS) into the 3D environment.

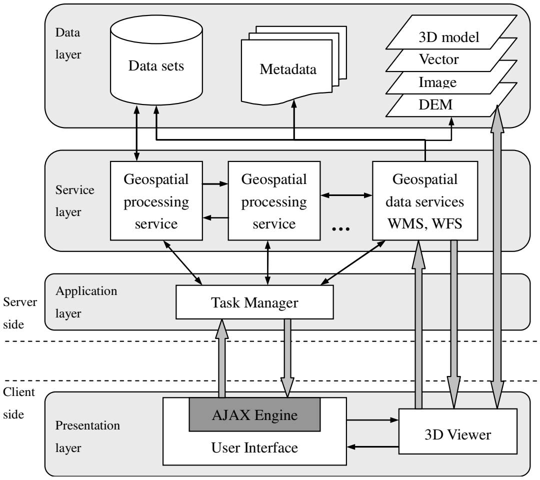


Fig. 1. The architecture and key components

3.2 The Application Layer

The Application layer addresses issues about service integration and chaining. As shown in Figure 1, the central component of this layer is the Task Manager which is a web application that can be built up by a general programming language (e.g. Java). The Task Manager has the capabilities of a general web application: it can handle the requests from the Presentation layer and send the processing results back to the client. What makes Task Manager different from other web applications is that it can organize several processing services into a workflow (a sequence of operations [19]) that can fulfill user's specific requirements.

3.3 The Service Layer and the Data Layer

The Service layer consists of distributed geospatial data services and processing services. The data services, such as WMS and WFS, provide the Task Manager and the 3D Viewer with standard interfaces to access the geospatial data deposited in the Data layer. The processing services deal with business logic, and can manipulate the datasets from either the data repositories or the geospatial data services. A processing service is also able to interact with other processing services in order to accomplish complex geospatial tasks. The Data layer provides the architecture with geospatial data which include 3d models, DEM, vector data, metadata, and other datasets.

3.4 Workflows of the Architecture

Two different workflows (shown in Figure 2) are designed in our architecture to respectively deal with tasks of different complexities.

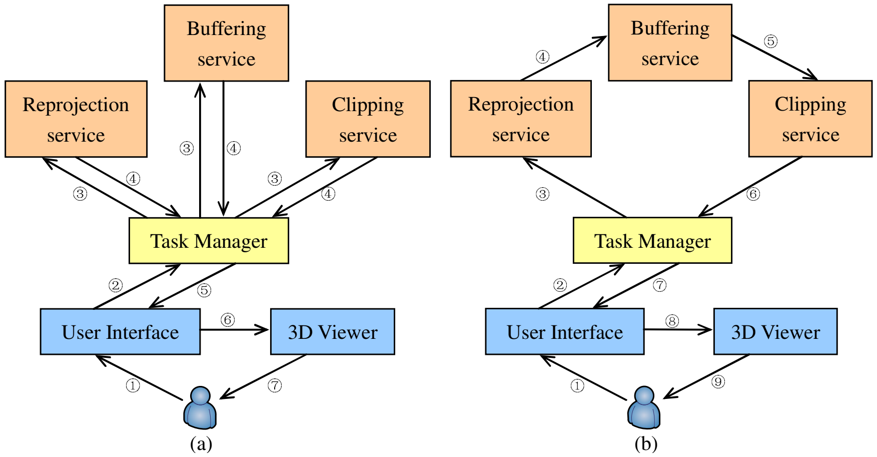


Fig. 2. The workflows of the architecture: (a) invoking a single service; (b) invoking a service chain

The workflow in Figure 2a is implemented when user's task can be easily handled by a single geospatial processing service (e.g. buffering service). User first inputs the request parameters to the User Interface which then sends user's request to the Task Manager. The Task Manager analyzes user's request and redirects such request to a corresponding processing service. The processing service executes its function and sends the results back to the Task Manager. The results are then passed to the User interface and the 3D Viewer in turn. Finally, the 3D Viewer visualizes the processing results in the virtual environment.

The workflow in Figure 2b is implemented, when user's task can only be accomplished by the cooperation of several processing services. In this workflow, when the Task Manager intercepts a new request, it does not immediately redirect this request to a particular service. Instead, it analyzes the request, identifies the services needed to accomplish this task, and organizes these services in a sequence. The Task Manager then sends the request to the first service which then passes its processing results to the next services. When the last processing service finishes its own work, it transmits the final results to the Task Manager. Finally, the 3D Viewer obtains the processing results and displays these results to end user.

4 Implementation

Our approach has been implemented in a 3D GIS project named Digital Chongming Island, Shanghai, China. DCI is a virtual globe-based web GIS and its 3D GIS platform is Skyline Globe. Our goal is to integrate geospatial processing services into DCI by using the approach proposed in this paper.

4.1 Preparing Geospatial Processing Service

The geospatial analysis functions must first be published as web services. While some of these functions (e.g. buffering) can be directly published by utilizing ArcGIS Server, other functions (e.g. the function of regional pollution evaluation which is based on an algorithm designed by the local government department) need to be developed and then published as services by Apache Axis. We also employ ArcGIS Server to publish the geospatial data of DCI as standard services (WMS and WFS) in order to make these heterogeneous data accessible to the processing services.

4.2 Integrating Geospatial Processing Services

By adopting our method, we have successfully integrated a variety of geospatial processing services into DCI. Since similar ways are used to combine the services and the system, we only describe the integration process of two services: the Routing service which is a simple service, and the SO_2 Concentration Distribution service which is a service chain.

4.2.1 The Routing Service

The Routing service can calculate the shortest path between two locations. This service relies on Chongming's road data which is published as WFS. To invoke this service, user first selects the start point and end point in the 3D environment. Then the coordinates of the two points along with a tag, which identifies the specific service requested by end user, are sent to the Task Manager. The Task Manager checks this tag and redirects these coordinates to the Routing service. In the next step, the Routing service works out the shortest path by referring to the road data and returns the vertexes of the path to the Task Manger. These vertexes are then passed to the 3D Viewer (the ActiveX control of Skyline Globe), and the Application Programming Interfaces (APIs) of Skyline Globe are employed to draw a route in the 3D environment according to these vertexes, and we also create a vehicle to vividly show the path to end user (Figure 3a).

4.2.2 The SO_2 Concentration Distribution Service

The SO_2 Concentration Distribution service shows the current distribution of the SO2 concentration in Chongming's local atmosphere. This service is a service chain which consists of three simple services: the SO_2 Monitoring service, the Point Layer Creating service, and the Spatial Interpolation service. The SO_2 Monitoring service retrieves the current monitoring data of SO_2 from the 102 air monitoring stations distributed on Chongming Island. The Point Layer Creating service can create a point layer based on the coordinates of points, and the Spatial Interpolation service is able to produce a surface image from a point layer.

To invoke the SO_2 Concentration Distribution service, user first send a request to the Task Manager, which then analyzes this request, organizes the three simple services and redirects the request to the SO_2 Monitoring service. In the next step, the SO_2 Monitoring service obtains the current monitoring data, and transmits them to the Point Layer Creating service, which creates a point layer from these data and passes this layer to the Spatial Interpolation service. The Spatial Interpolation service then

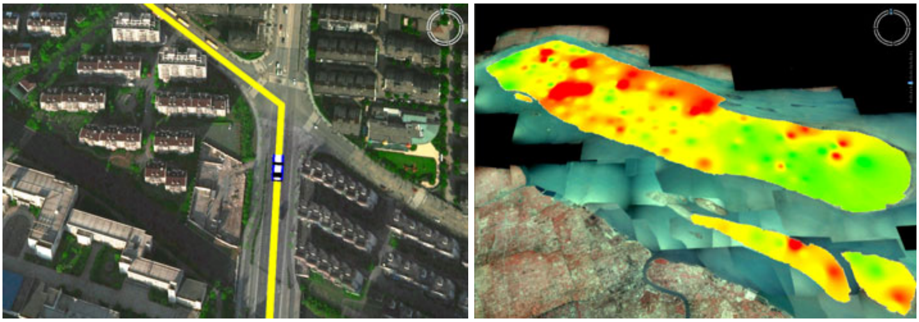


Fig. 3. Results of invoking geospatial processing services in DCI: (a) the Routing service; (b) the SO2 Concentration Distribution service

produces an image from this point layer and sends the URL of this image to the Task Manger. Finally, the 3D Viewer receives the URL of the image, and displays this image into the 3D environment (Figure 3b).

5 Conclusions

3D GIS excels in data visualization, but it is comparatively weak in geospatial analysis. This paper presents an approach for integrating geospatial processing services into 3D GIS in order to enhance its geospatial analysis capability. The architecture of our approach contains four layers which are the Presentation layer, the Application layer, the Service layer and the Data layer. We have also designed two different workflows to deal with tasks that have different complexities. While the first workflow involves only one simple geospatial processing service, the second workflow chains two or more services together.

Our approach has been implemented in a 3D GIS project named Digital Chongming Island, Shanghai, China. We publish geospatial analysis functions as web services, and then integrate these services into DCI by using the approach proposed in this paper. In particular, we have described the details of invoking two geospatial processing services in DCI. The results of the implementation have demonstrated that our approach can effectively combine 3D GIS and geospatial processing services.

Acknowledgments. The authors would like to thank the DCI partners for their valuable contributions to this research. DCI is a virtual globe-based decision supporting system which is sponsored by the Science and Technology Commission of Shanghai Municipality (Grant No. 07DZ12037).

References

1. Tuttle, B.T., Anderson, S., Huff, R.: Virtual globes: an overview of their history, uses, and future challenges. Geography Compass 2(5), 1478–1505 (2008)
2. Nature: Think global. Nature, vol. 439 (7078), p. 763 (2006)

3. James, D.M., Phillip, A.G.: A GIS-based borehole data management and 3D visualization system. Computers & Geosciences 32, 1699–1708 (2006)
4. Kreuseler, M.: Visualization of geographically related multidimensional data in virtual 3D scenes. Computers & Geosciences 26, 101–108 (2000)
5. Skyline Software System Inc.: SkylineGlobe Technology Overview, pp. 1–17 (2009)
6. Google Inc.: Google Earth User Guide, pp. 1–131 (2007)
7. World Wind Web Manual, `http://worldwind.arc.nasa.gov/manual.html`
8. Alonso, G., Casati, F., Harumi, K., Machiraju, V.: Web Services. Concepts, Architectures and Applications. Springer, Heidelberg (2004)
9. Foster, I.: Service-oriented science. Science 308(5723), 814–817 (2005)
10. Alameh, N.: Chaining geographic information web services. IEEE Internet Computing 7(5), 22–29 (2003)
11. Granell, C., Diaz, L., Gould, M.: Managing earth observation data with distributed geoprocessing services. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2007), pp. 4777–4780. IEEE CS, Los Alamitos (2007)
12. OGC: OpenGIS Service Architecture. OpenGIS Abstract Specification, 2–112 (2002)
13. Alameh, N.: Scalable and Extensible Infrastructures for Distributing Interoperable Geographic Information Services on the Internet. doctoral dissertation. MIT Libraries, Cambridge (2001)
14. Paul, M., Ghosh, S.K., Acharya, P.S.: Enterprise geographic information system (E-GIS): A service-based architecture for geo-spatial data interoperability. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE GRSS Press, Denver (2006)
15. Paul, M., Ghosh, S.K.: A service-oriented approach for integrating heterogeneous spatial data sources realization of a virtual geo-data repository. International Journal of Cooperative Information Systems 17(1), 111–153 (2008)
16. Granell, C., Diaz, L., Gould, M.: Service-oriented applications for environmental models: Reusable geospatial services. Environmental Modelling & Software 25, 182–198 (2010)
17. Dunne, D., Sutton, G.: 3D Web mapping: integrating marine data into NASA World Wind. Hydro International 10(9), 7–9 (2006)
18. Craglia, M., Goodchild, M.F., Annoni, A., Camara, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S., Parsons, E.: Next-Generation Digital Earth. A position paper from the Vespucci initiative for the advancement of geographic information science. International Journal of Spatial Data Infrastructures Research 3, 146–167 (2008)
19. Aalst, W., Hee, K.: Workflow Management: Models, Methods, and Systems. MIT Press, Cambridge (2002)

Parallel K-Means Clustering of Remote Sensing Images Based on MapReduce

Zhenhua Lv[1], Yingjie Hu[1], Haidong Zhong[1], Jianping Wu[1,*],
Bo Li[2], and Hui Zhao[2]

[1] Key Laboratory of Geographic Information Science, Ministry of Education;
Geography Department; East China Normal University, Shanghai, China
[2] Institute of Software Engineering, East China Normal University, Shanghai, China
zhenhualv@gmail.com,
jpwu@geo.ecnu.edu.cn

Abstract. The K-Means clustering is a basic method in analyzing RS (remote sensing) images, which generates a direct overview of objects. Usually, such work can be done by some software (e.g. ENVI, ERDAS IMAGINE) in personal computers. However, for PCs, the limitation of hardware resources and the tolerance of time consuming present a bottleneck in processing a large amount of RS images. The techniques of parallel computing and distributed systems are no doubt the suitable choices. Different with traditional ways, in this paper we try to parallel this algorithm on Hadoop, an open source system that implements the MapReduce programming model. The paper firstly describes the color representation of RS images, which means pixels need to be translated into a particular color space CIELAB that is more suitable for distinguishing colors. It also gives an overview of traditional K-Means. Then the programming model MapReduce and a platform Hadoop are briefly introduced. This model requires customized 'map/reduce' functions, allowing users to parallel processing in two stages. In addition, the paper detail map and reduce functions by pseudo-codes, and the reports of performance based on the experiments are given. The paper shows that results are acceptable and may also inspire some other approaches of tackling similar problems within the field of remote sensing applications.

Keywords: K-Means; remote sensing; parallel; MapReduce; Hadoop.

1 Introduction

The K-Means clustering is one of the most common methods of data analysis, as in the field of pattern recognition, data mining, image processing, etc. It is useful in the field of remote sensing analysis as well, where objects with similar spectrum values are clustered together without any former knowledge. It has become the basic algorithm of unsupervised classification, providing us an overview of objects easily and directly, with which our further analysis becomes clear. The time complexity of

[*] Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 162–170, 2010.

K-Means, however, is considerable, and the execution is time-consuming and memory-consuming especially when both the size of input images and the number of expected classifications are large. To improve the efficiency of this algorithm, many variants have been developed. It is commonly believed that there are two ways to reduce the time consumption, the first is concerned with optimizing the algorithm itself, whereas another one focuses on changing the proceeding of execution, that is migrate the sequential process to parallel environment. While in our opinion, optimization of sequential K-Means algorithm is important and has made much great success, this paper prefers to have the algorithm running under parallel environment, which will be considered as the appropriate way to process large amounts of data set. Different with traditional methods that implemented based on MPI [1] [2], we use MapReduce as our basic computing model, which is first proposed by Google Corporation in 2004 and has now been widely welcomed in many domains. Related work has been done, however they both have two faults that will be discussed in the third section.

The rest of this paper is organized as follows. In the next section we will briefly describe the traditional approach that clustering RS images sequentially. It includes color space we used and the clustering method K-Means. Then, at the third section we introduce the related work and discuss our implementation. In the fourth section, experiments are conducted. Finally, merits and demerits of this paper are pointed out.

2 Sequential Process

The input of clustering is a color remote sensing image, a set of pixels each of which represented by RGB value. We expect the output, after the K-Means algorithm, is k subsets of pixels, within each subset the pixels have the most similar color. The similarity in color is better understood in a transformed space called CIELAB rather than RGB. So the preparation for clustering is a transformation of each pixel from RGB-value to Lab-value, and the similarity of pixels is represented by the distance between Lab-values.

2.1 Transformation of Color

The L*a*b* color space (also called CIELAB) is derived from the CIE XYZ tri-stimulus values, and it consists of a luminosity 'L*' or brightness layer, chromaticity layer 'a*' indicating where color falls along the red-green axis, and chromaticity layer 'b*' indicating where the color falls along the blue-yellow axis [3]. This color space describes all colors visible to the human eye and is created to serve as a device-independent model. The formula for converting RGB into LAB is given by [4], and this conversion program implemented in MATLAB can be freely downloaded on the web [5]. In this paper, we have made use of another free function for color conversion that was written in Java and available on [6].

We use P(L, a, b) representing a pixel value in LAB color space. Since the color information exists in ‘a*b*’ space, we will measure the difference of two colors by ‘a*’ and ‘b*’ values, ignoring ‘L*’ values. The difference can be defined by Euclidean distance. The greater the distance of two pixels, the less similar they are or vice versa.

2.2 Overview of K-Means Clustering

Let P_1, P_2, ..., P_n be the set of N pixels, where P_i is the i^{th} pixel consisting of a (a, b) pair. In the pair, a, b are color components in L*a*b* color space. Both are calculated from RGB values the way we have discussed in previous section. Since the pixels in image files are arranged along widths and heights like rectangular shapes and each pixel has a two dimensions coordinates, a conversion of pixels to sequential pairs is required. Let C_1, C_2, ..., C_k be the K clusters, where K is a input parameters for the algorithm. Let m_1, m_2, ..., m_k be the centroids associated with their clusters, such that mi is the centroid of cluster C_i , for $1 \le i \le K$.

$$J_e = \sum_{j=1}^{K} \sum_{P_i \in C_i} \left\| P_i - m_i \right\|^2 \tag{1}$$

$$m'_i = m_i + \frac{1}{N_i - 1}[m_i - P_i] \tag{2}$$

$$m'_k = m_k + \frac{1}{N_k + 1}[P_i - m_k] \tag{3}$$

The quantity of clustering is measured by the error variance defined as (1) [7]. Ideally, the optimized solution of clustering is the one that makes J_e reaching the minimum. When P_i has been moved out its cluster C_i, there causes a decrease in the centroid m_i, generating the new centroid m'_i as (2). Similarly, when P_i has been moved into another cluster C_k, there cause an increase in the centroid m_k, generating the new centroid m'_k as (3).

Our work in this paper considers the K-Means algorithm executing as the following steps:

Step 1. Generate an initial K clusters.

Step 2. Move pixels between clusters when the sum of error variance is going to decrease. Select each pixel P, then let C_i be P's cluster, N_i be the number of pixels in this cluster and m_i be the centroid. If N_i =1, then ignore this pixel and select another one, otherwise calculate the variation by (4):

$$v_j = \begin{cases} \dfrac{N_j}{N_j + 1} \left\| P - m_j \right\|^2 \; j \ne i \\ \dfrac{N_i}{N_i - 1} \left\| P - m_i \right\|^2 \; j = i \end{cases} \tag{4}$$

In (4), i is the number indicating the cluster P belongs to, while j varies from 1 to K. v_j (j=i) is a decrease in the sum of error variance when P has be moved out of its cluster C_i. v_j becomes an increase in the sum of error variance when $j \ne i$ and on the assumption that P has be assigned to cluster C_j. If there exist a $j(j \ne i)$

satisfying $v_j < v_i$, that means moving P from C_i to C_j gives rise to a decrease in sum of total error variance, which is the purpose of this step, otherwise, jump to the beginning of this step. If there existing several j satisfying such condition, the one with the minimal increase would be chosen.

Step 3. Move the pixel from C_i to C_j and update m_i to m_j.

Step 4. Iterates this procedure for N times.

This algorithm doesn't stop until there is no variation of J_e took place or maximum of iteration specified by implementation is reached.

3 Parallel Process

MapReduce [8] is a programming model and Hadoop [9] is an open source system that implements the MapReduce model. The basic primitives of this model are map/reduce functions both are written by users.

3.1 Related Work

In paper [10], algorithm is divided into four steps briefly as follows.

Step 1: generate initial guess.

Step 2: Map: (null, data-instance) -> list (cluster-id, (data-instance, 1)), where the in-key is null, and the in-value is the data instance vector.

Step 3: Reduce: (cluster-id, (data-instance, count)) -> list (cluster-id, (sum-of-data-instances, number-of-instances))

In the output of reduce, the sum-of-data-instances divided by number-of-instances is the mean of the cluster. Simply put, their methods can be summarized as:

1. Random generate k points as initial k means.
2. Apply K-Means Map & Reduce.
3. While not done, go to Step 2.

The idea of paper [11] resembles the previous one. A slight difference is they add a combiner function to improve the performance. However, both the methods may have two faults in the view of processing RS images. For one thing, the final output is only the centroid representing the cluster, which is useless for RS classification. Instead, the output we expect should indicate the cluster of each pixel, such as appending a cluster identifier to each pixel instance. For another, dynamically updating the centroid is missing in their methods, and error variances are not considered either. Therefore, the final output will largely depend on the initial centroid specified by the user, giving rise to great deviations if inappropriate centroid is selected at the beginning.

3.2 Implements on Hadoop

In the parallel environment, dynamically updating centroid in each step needs global communication between computer nodes, which should be avoided. Hence, in our method pixels are clustered partially. In general, there are two Hadoop jobs for calculations, one is to distribute each pixel into its initial cluster, and another is to move pixels minimizing the error variances.

To simply suit the framework, input images are translated into text file. Each line represents a pixel like (file_id, pixel_id, r, g, b), where file_id identifies the file, pixel_id is the line number in each file, and r, g, b are the values.

Our algorithm can be detailed as two steps. Each of them is represented by a job of Hadoop. The first step is to initial clustering, in which pixels closest to the centroid are clustered. To identify the owner cluster of a pixel, we insert the cluster number to a text line that represents the pixel. A job of the Hadoop has three functions, they are configure, map and close. In the configure function, the job reads some necessary variables from the JobConf object that are distributed by the name node. The variables are initial centroids specified by the user's description file. The map function performs chief calculations of clustering. The close function saves the final centroids to a file (called centroid file). Since there are many map tasks running on different computer nodes when the job is committed to Hadoop, the input pixels of one map task are only a subset. There are many centroid files, and they should be merged for the next step. The method that merges centroid files here is to simply take the mean of each cluster centroid as the new centroid.

The second step is to minimize the error variances. The method how to minimize the error variances is mentioned in the previous section and it is abided by (6). Reduce functions in both the jobs do no extra work but write the final results to files (called result file). Thus the results files contain cluster values indicating each pixel's cluster when all the jobs are accomplished.

It needs two job client objects in Hadoop to implement these two steps. The output of the first job is in turn regarded as the input of the second job. The two jobs run step by step.

Step 1. Initial clustering. The pseudo-codes describe the job 'Initial clustering' as shown in Table 1.

In 'Initial clustering', DETA is the increase of centroid when a new pixel is moved in its cluster. It can be calculated by the expression (3). The input 'CountsInCluster' is a counter list in which each item represents the number of pixels that a cluster contains. It is initialized at the configure function. The 'lab_distance' method is to calculate Euclidean distance of two pixels.

Step 2. Minimize the total error variance for each cluster. This job reads its input from the files that are produced by the first job. At the configure function of this job, it merges some partial results. For example, centroids of each cluster which are produced by each map task are collected together. And some 'CountsInCluster' values are also added together. The map function of the job 'Minimize Error' is shown in Table 2. The idea is to move the pixel to another cluster if the sum of error variance decreases. The error variance of a cluster will decrease when a pixel of it moves out and will increase then a pixel moves in. If the amount of increase is less than that of decrease, the pixel is of course worth moving.

Table 1. Initial Clustering

```
InitalClusteringJob.map

Input : (in_useless, in_pixel), where ‘in_pixel’ is (file_id, pixel_id, r, g, b);
Centroid{K} = getFromConfiguration(); // assign K initial centroid
CountsInCluster{K} = {1};// each cluster has 1 pixel at the beginning
Output : (out_cluster_id, out_ pixel)
   rgbPixel = parseFrom(in_pixel);        labPixel = Transform(rgbPixel);
   minDistance = Double.MAX_VALUE;   closestClusterId = -1;
   for (i = 0; i < K; ++i) {
      dist = lab_distance(input, Centroid[i]);
      if (dist < minDistance) { minDistance = dist;    closestClusterId = i; }
   }
   // move pixel into the closest cluster and update the centroid
   ++ CountsInCluster [closestClusterId];
   Centroid [closestClusterId] = Centroid [cid] + DETA;
   out_cluster_id = closestClusterId;
   output = {closestClusterId, file_id, pixel_id, labPix}
```

Table 2. Minimize Error

```
MinimizeErrorJob.map

Input : (cluster_id, labPixel)
Centroid{K}; //a set of centroids
Counts{K}; //Counts[i] indicates the count of pixels in cluster i
Output : (cluster_id, out_ pixel)
   // move the pixel to another cluster if error variance decreases
   icount =  Counts [cluster_id];    idestCluster = -1;
   cen = Centroid [cluster_id];  double tempMin = decrease;
   decrease = icount / (icount - 1) * lab_distance(labPixel, cen);
   for (int k = 0; k < ClusterNumber; ++k) {
       if (k != cluster_id) {
          increase = icount / (icount+1) * lab_distance(input, cen);
       if (increase < tempMin) { idestCluster = k; tempMin = increase;} } }
    if ( idestCluster != -1) {
      cen = cen + (cen - labPixel) / (icount - 1);
      cen2 = Centroid[idestCluster];
      cen2 = cen2 + (labPixel – cen2) / (icount + 1);
      Centroid[cluster_id] = cen;
      Centroid[idestCluster] = cen2;
      --m_Counts[cluster_id]; ++m_Counts[idestCluster];
     output = { idestCluster, labPixel}  }
```

The K-Means algorithm requires that iterations for minimizing error variances run more than one times and select the best result. In our algorithm the iteration runs only once, because we found the first iteration decreasing the error variances immensely and the consequent iterations take trivial effects.

4 Experiments

The hardware we used was a computer cluster which had one control node and eight computing nodes. The control node was composed of eight 2.0GHz-CPUs and 4GB memory, while each computing node was composed of eight 2.0GHz-CPUs and 8GB memory. All nodes were connected by a gigabit switch. The total capacity of the cluster is about 20TB.

We implemented sequential K-Means algorithm by using C++ and it ran well if the size of the input image is not large. In the experiment of sequential algorithm the size cannot exceed 1000 pixels both in width and height. We also implemented another version called parallel K-Means running on Hadoop in Java. The image used here contained vegetations and constructions as Fig. 1 (a) shown. The two images after clustering were divided as Fig. 1 (b) and (c) shown. The same image was also taken as the input of parallel K-Means on Hadoop, and the results were shown in Fig. 2 (b) and (c) separately. It is obvious that the results on different environments are very similar. Their corresponding error variances of these two algorithms were summarized, as Table 3 shown. The column deviation was calculated by taking the results of sequential K-Means as criterion.

Another kind of images was performed by the parallel K-Means, but we cannot visualize them because their sizes and amounts were very large. They cannot be examined by the sequential algorithm either. The data were aerial remote sensing images about some districts of Shanghai in China. The data were composed of many blocks and seemed like lattice, each of which was an image with tif format. Each image was 4000 pixels in width and 3200 pixels in height. The size of each image file was about 38 megabytes. But the amount of the file size soars to about 240 megabytes when the form of the pixels was changed into text. It was unwise to change the form, but there was no way to directly operate tif-format images on Hadoop. As Table 4 shown, the cost of time increased dramatically when the count of images became larger. This might be caused by the data transmission of network.

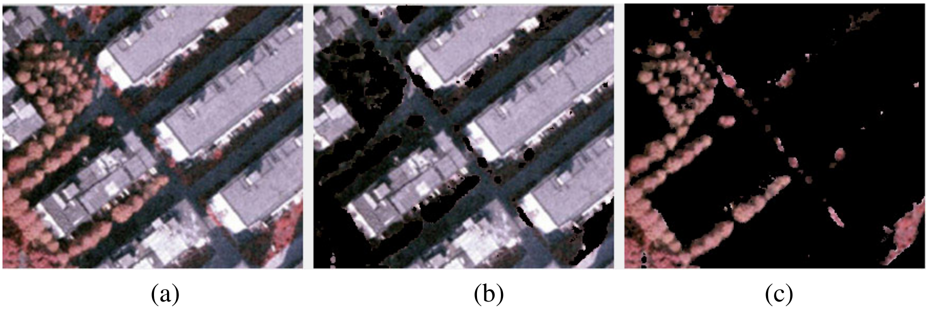


Fig. 1. Results of sequential K-Means, where K is two. (a) is the original image. (b) is one of its clusters and (c) is another cluster.

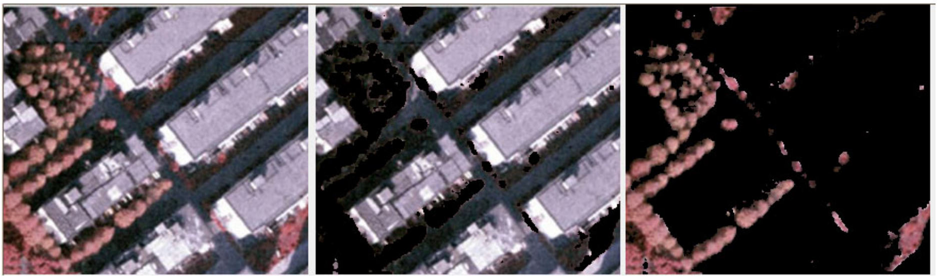

(a) (b) (c)

Fig. 2. Results of parallel K-Means on Hadoop, where K is two. (a) is the original image. (b) is one of its cluster and (c) is another cluster.

Table 3. Deviations of error variances

Cluster Name	Error Variance of Sequential K-Means	Error Variance of Parallel K-Means	Deviation
Vegetation	1013060.91	946741.75	-6.55%
Construction	3040686.81	3104197.87	2.09%

Table 4. Time cost

Image count	Size in pixels (width * height)	File size in bytes	Time cost (second)
1	4,000*3,200	248M	36
4	8,000*6,400	995M	125
20	20,000*12,800	4.85G	778
80	40,000*25,600	19.5G	5413

5 Conclusion and Future Work

In this paper, the pixels are firstly clustered in the partial scope, and then they are collected together and merged. In the view of mathematics, this slightly deviates from the origin of the algorithm. But, with the amount of input increasing, the deviations of this method slow down and can satisfy the concrete work. Since the Hadoop APIs have not supported to handle images as input directly by now, there need a transformation for images from binary into text files. This inevitably enlarges the size of input data and reduces the performance. In the near future it may provide a direct way processing images on Hadoop.

In fact, the final purpose of clustering RS images is to picture an overview of the data and provide us with approximate figures for further research. There is no technique that performs clustering RS images completely satisfying the need of work without artificial participation. Hence, instead of evaluating the results by error variances or other methods, visualizing the output and judging by human eyes directly are more practical. We have developed a simple tool for visualizing, but it is not suitable for handling large images. And this is another point of improvement for our work.

References

1. Gursoy, A.: Data Decomposition for Parallel K-means Clustering. Parallel Processing and Applied Mathematics 14, 241–248 (2004)
2. Li-shun, J., Ding-sheng, L.: Research on K-Means Clustering Parallel Algorithm of Remote Sensing Image. Remote Sensing Information 01, 27–30 (2008)
3. Matlab R2008a Product Help. Demos/Toolboxes/Image Processing/Image Segmentation/ Color-Based Segmentation Using the L*a*b* Color Space
4. Kartikeyan, B., Sarkar, A., et al.: A segmentation approach to classification of remote sensing imagery. International Journal of Remote Sensing 19, 1695–1709 (1998)
5. MATLAB Central - File detail - RGB2Lab, `http://www.mathworks.com/matlabcentral/fileexchange/24009`
6. Color Inspector 3D - Color Space Conversions, `http://www.f4.fhtw-berlin.de/~barthel/ImageJ/ColorInspector/HTMLHelp/farbraumJava.htm#rgb2lab`
7. Zhaoqi, B., Xuegong, Z.: Pattern Recognition, 2nd edn., pp. 235–237. Tsinghua University Press, Beijing (2000)
8. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Communications of the ACM 51, 107–113 (2008)
9. Hadoop Website, `http://hadoop.apache.org/`
10. Yao, K.T., Lucas, R.F., et al.: Data Analysis for Massively Distributed Simulations. Interservice/Industry Training, Simulation, and Education Conference(I/ITSEC) (Got from Google Scholar) (2009)
11. Zhao, W., Ma, H., et al.: Parallel K-Means Clustering Based on MapReduce. In: CloudCom 2009. LNCS, vol. 5931, pp. 674–679 (2009)

Enhancing Distributed Web Security Based on Kerberos Authentication Service

Cao Lai-Cheng[*]

School of Computer and Communication,
Lanzhou University of Technology, Lanzhou 730050, China
caolch@lut.cn

Abstract. The increasing popularity of distributed web has promoted the development of new techniques to support various kinds of applications. However, users are faced with insecurity due to its inherent untrustworthiness. An identity (ID) authentication mechanism was presented. Using Kerberos protocol, Local web and Remote web could authenticate the client. If mutual authentication was required, client could also authenticate Local web and Remote web. Moreover, encryption function in the authentication process adopted Rijndael encryption algorithm of AES (Advanced Encryption Standard). Security analysis proves that this authentication process is no-impersonating and has highly availability, and also shows it is transparent and scalable and resisting attack.

Keywords: distributed web security; Kerberos protocol; authentication; Rijndael encryption algorithm.

1 Introduction

The distributed web has experienced tremendous growth in connecting millions of computers all over the world, and has made it possible for users to share resources across the Internet. As connectivity and sharing increase, security and privacy issues are becoming increasingly critical. The attacks against the distributed web have various resources [1, 2], including the web servers, the communication links, the authentication and authorization mechanisms, the data connectivity between the application and the backend databases, etc. In contrast to traditional methods for securing closed systems, we need a new strategy for protecting our web in a distributed environment. Sascha Rehbock [3] proposes solutions based on authentication standards for enabling TNC (Trusted Network Connect) in open, web-based scenarios, this approach provides assurance as to the security state of clients accessing security sensitive web-based services. Alessandro Basso [4] presents MosaHIP, a Mosaic-based Human Interactive Proof (HIP), which is able to prevent massive automated access to web resources. Hany F. EL Yamany [5] applies three different mining techniques based on the

[*] This work is supported by the National Natural Science Foundation of China under Grant No. 60972078; the Gansu Provincial Natural Science Foundation of China under Grant No. 0916RJZA015.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 171–178, 2010.

Association Rules to help predicting attacks. Song Han [6] uses three-party key establishment to enable secure communications for Service Requester and Service Provider through web services. But the above methods cannot well provide authentication function, assume an open distributed environment in which users at the web wish to access the web services on servers distributed throughout the network. Kerberos is a trusted third-party authentication service. It provides a centralized authentication server whose function is to authenticate users to servers and servers to users. Thus, we use Kerberos protocol to construct our distributed web; it can provide authentication functions that we want.

This paper begins with an introduction. In section II we describe our security architecture design. In Section III the authentication process of the distributed web is presented. In section IV we finish security analysis about our method. The conclusion is presented in section V.

2 Security Architecture Design

In the distributed environment, an unauthorized user may be able to gain access to the web and data that he or she is not authorized to access. In order to protect user information and web resources, we require that client systems authenticate themselves to the web, but trust the client system concerning the identity of its user. We design the security architecture of distributed web by Kerberos protocol [7, 8] (as shown in Fig. 1), it includes two representative realms, namely Realm A and Realm B. Realm A includes Client, Switcher, Authentication Server (*AS*), Ticket-granting Server (*TGS*), Local web. Realm B includes Switcher, *AS*, *TGS*, Remote web. Authentication Server keeps a database containing the private keys of the clients and all of servers. Realm A and Realm B are connected to Internet with router; the client and the web (Server) can communicate each other based on TCP/IP protocol.

- *Realm*: Indicates realm of the client.
- *Client*: Requires to gain access to Local web in Realm A or Remote web in Realm B.
- *AS*: Authenticates servers to the client.
- *TGS*: Grants service-granting ticket to the client.
- *Local web*: Stores resource and data for the local uses to share directly in same realm.
- *Remote web*: Stores resource and data for the remote uses to share in another realm.

3 Authentication Process

3.1 Local Authentication

If the client requests to access Local web in same Realm, the authentication, which the server authenticates the client or the client authenticates the server, belongs to local authentication.

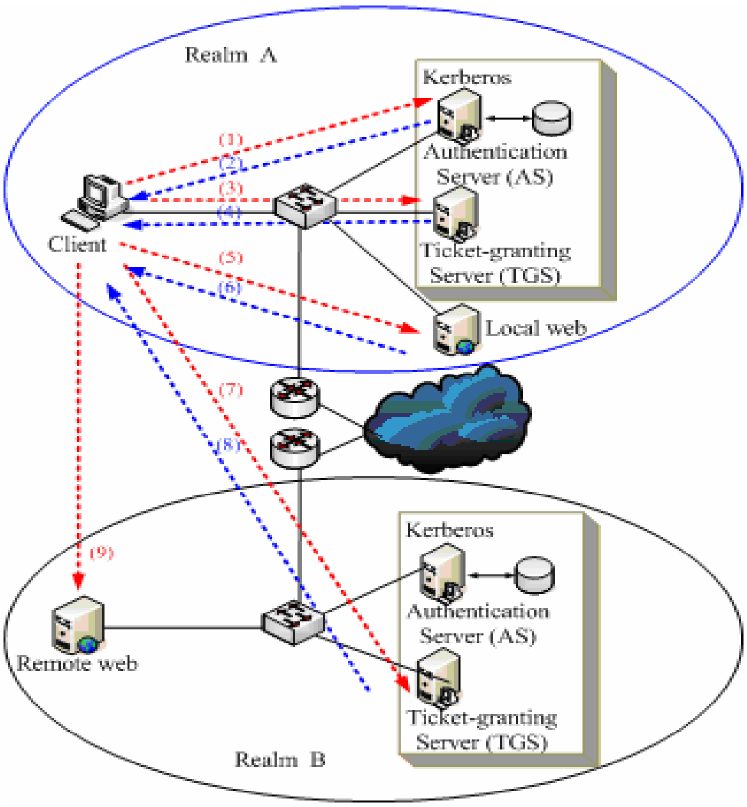

Realm A
Kerberos
Authentication Server (AS)
Ticket-granting Server (TGS)
Client
Local web
(1)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
Kerberos
Authentication Server (AS)
Ticket-granting Server (TGS)
Remote web
Realm B

Fig. 1. Security architecture about distributed web

1) Authentication process

- *Authentication service exchange to obtain ticket-granting ticket*

Message (1): Client requests ticket-granting ticket (corresponding to (1) in Fig. 1):

$$C \to AS : Options \parallel ID_c \parallel \mathrm{Re}\,alm_c \parallel ID_{tgs} \parallel Times \parallel Nonce_1$$

Message (2): *AS* returns ticket-granting ticket (corresponding to (2) in Fig. 1):

$$AS \to C : \mathrm{Re}\,alm_c \parallel ID_c \parallel Ticket_{tgs} \parallel E_{K_c}(K_{c,tgs} \parallel Times \parallel Nonce_1 \parallel \mathrm{Re}\,alm_{tgs} \parallel ID_{tgs})$$

Where, $Ticket_{tgs} = E_{K_{tgs}}(Flags \parallel K_{c,tgs} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)$, the symbols are showed in Table 1.

- *Ticket-granting service exchange to obtain service-granting ticket*

Message (3): Client requests service-granting ticket (corresponding to (3) in Fig. 1):

$$C \to TGS : Options \parallel ID_{lw} \parallel Times \parallel Nonce_2 \parallel Ticket_{tgs} \parallel Authentica\,tor_c$$

Where, $Authentica\,tor_c = E_{K_{c,tgs}}(ID_c \parallel \mathrm{Re}\,alm_c \parallel TS_1)$

Message (4): *TGS* returns service-granting ticket (corresponding to (4) in Fig. 1):

$$TGS \to C : \mathrm{Re}\,alm_c \parallel ID_c \parallel Ticket_{lw} \parallel E_{K_{c,tgs}}(K_{c,lw} \parallel Times \parallel Nonce_2 \parallel \mathrm{Re}\,alm_{lw} \parallel ID_{lw})$$

Where, $Ticket_{lw} = E_{K_{lw}}(Flags \parallel K_{c,lw} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)$

Where, the symbols are showed in Table 2.

Table 1. Symbols of message (1) and message (2)

Symbol	Meaning
C	Client
AS	Authentication server
Options	Used to request that certain flags be set in the returned ticket
ID_c	Tells AS identity of user from this client
$Realm_c$	Tells AS realm of user from this client
ID_{tgs}	Tells AS identity of TGS that user requests access
Times	Used by the client to request to the time settings in the ticket, it consists the desired start time, the requested expiration time and the requested renew expiration time
$Nonce_1$	A random value that client produces to be repeated in message (2) to assure that the response is fresh and has not been replayed by an attacker
$Ticket_{tgs}$	Ticket to be used by client to access TGS
K_c	Encryption is based on client's password, enabling AS and user to verify password, and protecting contents of message (2)
$K_{c,tgs}$	Copy of session key accessible to client created by AS to permit secure exchange between client and TGS without requiring to share a permanent key
$Realm_{tgs}$	Tells client realm of TGS
K_{tgs}	Ticket is encrypted with key known only to AS and TGS, to prevent tampering
AD_c	Prevents use of ticket from client other than one that initially requested the ticket
$E(\ )$	Encryption function based on Rijndael encryption algorithm of AES

Table 2. Symbols of message (3) and message (4)

Symbol	Meaning
ID_{lw}	Tells TGS identity of Local web
$Nonce_2$	A random value that client produces to be repeated in message (4) to assure that the response is fresh and has not been replayed by an attacker
$Authenticator_c$	Client transmits an authenticator, which includes the ID and address of client's user and a timestamp
$Ticket_{lw}$	Ticket to be used by client to access Local web
$K_{c,lw}$	Copy of session key accessible to client created by TGS to permit secure exchange between client and Local web without requiring to share a permanent key
K_{lw}	Ticket is encrypted with key known only to TGS and Local web, to prevent tampering
$Realm_{lw}$	Tells client realm of Local web
TS_1	Informs TGS of time this authenticator was generated

Table 3. Symbols of message (5) and message (6)

Symbol	Meaning
LW	Local web
TS_2	Informs Local web of time this authenticator was generated
Subkey	The client's choice for an encryption key to be used to protect this specific application session. If this field is omitted, the session key uses the ticket ($K_{c,w}$)
Seq#	An optional field that specifies the starting sequence number to be used by Local web for messages sent to the client during this session. Message may be sequence number to detect replays

- *Client/Local web authentication exchange to obtain service*

Message (5): Client requests service (corresponding to (5) in Fig. 1):

$$C \rightarrow LW\ : Options\ \| Ticket_{lw} \| Authenticator_c$$

Where, $Ticket_{lw} = E_{K_{lw}}(Flags \| K_{c,lw} \| \mathrm{Re}alm_c \| ID_c \| AD_c \| Times)$

$$Authenticator_c = E_{K_{c,lw}}(ID_c \| \mathrm{Re}alm_c \| TS_2 \| Subkey \| Seq\#)$$

Message (6): Optional authentication of Local web to client (corresponding to (6) in Fig. 1): $LW \rightarrow C : E_{K_{c,lw}}(TS_2 \parallel Subkey \parallel Seq\#)$

Where, the symbols are showed in Table 3.

2) Justification analysis

Message (1): Client sends a message to the *AS* requiring access to the TGS, the *TGS* can acquire client's realm, client's ID and TGS's ID.

Message (2): The *AS* responds with a message, which mainly includes $Ticket_{tgs}$ for accessing *TGS* and a cryptograph encrypted with the user's password (K_c) by Rijndael encryption algorithm. The cryptograph contains a copy of the session key, $K_{c,tgs}$, where the subscripts indicate that this is a shared key for *C* and *TGS*. Because $K_{c,tgs}$ is inside the cryptograph encrypted with K_c, user can decrypt this cryptograph to gain $K_{c,tgs}$ as follow:

$$D_{K_c}(E_{K_c}(K_{c,tgs} \parallel Times \parallel Nonce_1 \parallel \mathrm{Re}\,alm_{tgs} \parallel ID_{tgs})) = K_{c,tgs} \parallel Times \parallel Nonce_1 \parallel \mathrm{Re}\,alm_{tgs} \parallel ID_{tgs}$$

$K_{c,tgs}$ can be read only by the user, $K_{c,tgs}$ is also included in the $Ticket_{tgs}$ that can be read only by the TGS. Therefore, both *C* and the *TGS* securely deliver $K_{c,tgs}$. Times in message (1) and message (2) can make *C* and the *TGS* to know that the message is timely.

Message (3): Armed $Ticket_{tgs}$ with and $K_{c,tgs}$, *C* can approach the *TGS*. In addition, *C* transmits $Authenticator_c$, which includes the ID and network address of C's user. The *TGS* can decrypt $Ticket_{tgs}$ with K_{tgs}:

$$\begin{aligned} D_{K_{tgs}}(Ticket_{tgs}) &= D_{K_{tgs}}(E_{K_{tgs}}(Flags \parallel K_{c,tgs} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)) \\ &= Flags \parallel K_{c,tgs} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times \end{aligned}$$

The *TGS* obtains the session key $K_{c,tgs}$ shared with *C*, it uses this session key to decrypt $Authenticator_c$:

$$D_{K_{c,tgs}}(Authentica\ tor_c) = D_{K_{c,tgs}}(E_{K_{c,tgs}}(ID_c \parallel \mathrm{Re}\,alm_c \parallel TS_1)) = ID_c \parallel \mathrm{Re}alm_c \parallel TS_1$$

The *TGS* can then check the name and address from $Authenticator_c$ with that of the $Ticket_{tgs}$ and with the network address of the incoming message. If all match, then the *TGS* is assured that the sender of the $Ticket_{tgs}$ is indeed the $Ticket_{tgs}$'s real owner.

Message (4) follows the form of Message (2). The cryptograph contains a copy of the session key, $K_{c,lw}$. Because $K_{c,lw}$ is inside the cryptograph encrypted with $K_{c,tgs}$, user can decrypt this cryptograph to gain $K_{c,lw}$ as follow:

$$D_{K_{c,tgs}}(E_{K_{c,tgs}}(K_{c,lw} \parallel Times \parallel Nonce_2 \parallel \mathrm{Re}\,alm_{lw} \parallel ID_{lw})) = K_{c,lw} \parallel Times \parallel Nonce_2 \parallel \mathrm{Re}\,alm_{lw} \parallel ID_{lw}$$

$K_{c,lw}$ can be read only by the user, $K_{c,lw}$ is also included in the $Ticket_{lw}$ that can be read only by Local web. Therefore, both *C* and Local web securely deliver $K_{c,lw}$.

C now has a service-granting ticket for Local web. When *C* presents this ticket $Ticket_{lw}$ as shown in Message (5), in addition, it sends $Authenticator_c$. Local web can decrypt $Ticket_{lw}$ with K_{lw}:

$$\begin{aligned} D_{K_{lw}}(Ticket_{lw}) &= D_{K_{lw}}(E_{K_{lw}}(Flags \parallel K_{c,lw} \parallel \mathrm{Re}alm_c \parallel ID_c \parallel AD_c \parallel Times)) \\ &= Flags \parallel K_{c,lw} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times \end{aligned}$$

Local web obtains the session key $K_{c,lw}$ shared with *C*, it uses this session key to decrypt $Authenticator_c$:

$$\begin{aligned} D_{K_{c,tgs}}(Authenticator_c) &= D_{K_{c,tgs}}(E_{K_{c,lw}}(ID_c \parallel \mathrm{Re}alm_c \parallel TS_2 \parallel Subkey \parallel Seq\#) \\ &= ID_c \parallel \mathrm{Re}\,alm_c \parallel TS_2 \parallel Subkey \parallel Seq\# \end{aligned}$$

Local web can then check the name and address from $Authenticator_c$ with that of the $Ticket_{lw}$ and with the network address of the incoming message. If all match, then Local web is assured that the sender of the $Ticket_{lw}$ is indeed the $Ticket_{lw}$'s real owner. Local web can also gain *Subkey*, and uses it to protect succeeding application session.

If mutual authentication is required, Local web can reply as shown in Message (6), *C* can use the key $K_{c,lw}$ to decrypt this message:

$$D_{K_{c,lw}}(E_{K_{c,lw}}(TS_2 \parallel Subkey \parallel Seq\#)) = TS_2 \parallel Subkey \parallel Seq\#$$

The *C* checks the TS_2 and *Subkey* and *Seq#* from the decrypted message. If all match ones in $Authenticator_c$ of the message (5), then the *C* can assure the ID of Local web because of only Local web can decrypt message (5) with its key K_{lw}.

3.2 Remote Authentication

If the client requests to access Remote web in different Realm, the authentication belongs to remote authentication.

1) Authentication process

Message (1): Client requests ticket-granting ticket for local *TGS* (corresponding to (1) in Fig. 1): $C \rightarrow AS : Options \parallel ID_c \parallel \mathrm{Re}\,alm_c \parallel ID_{tgs} \parallel Times \parallel Nonce_1$

Message (2): *AS* returns ticket-granting ticket for local *TGS* (corresponding to (2) in Fig. 1): $AS \rightarrow C : \mathrm{Re}\,alm_c \parallel ID_c \parallel Ticket_{tgs} \parallel E_{K_c}(K_{c,tgs} \parallel Times \parallel Nonce_1 \parallel \mathrm{Re}\,alm_{tgs} \parallel ID_{tgs})$

Where, $Ticket_{tgs} = E_{K_{tgs}}(Flags \parallel K_{c,tgs} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)$

Message (3): Client requests ticket-granting ticket for remote *TGS* (corresponding to (3) in Fig. 1): $C \rightarrow TGS : Options \parallel ID_{rtgs} \parallel Times \parallel Nonce_2 \parallel Ticket_{tgs} \parallel Authentica\ tor_c$

Where, $Authentica\ tor_c = E_{K_{c,tgs}}(ID_c \parallel \mathrm{Re}\,alm_c \parallel TS_1)$, ID_{rtgs} denotes identity of remote *TGS* (*rtgs* or *RTGS*).

Message (4): Local *TGS* returns ticket-granting ticket for remote *TGS* (corresponding to (4) in Fig. 1):

$$TGS \rightarrow C: \mathrm{Re}\,alm_c \parallel ID_c \parallel Ticket_{rtgs} \parallel E_{K_{c,tgs}}(K_{c,rtgs} \parallel Times \parallel Nonce_2 \parallel \mathrm{Re}\,alm_{rtgs} \parallel ID_{rtgs})$$

Where, $Ticket_{rtgs} = E_{K_{rtgs}}(Flags \parallel K_{c,rtgs} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)$

Message (7): Client requests ticket-granting ticket for Remote web (corresponding to (7) in Fig. 1): $C \rightarrow RTGS : Options \parallel ID_{rw} \parallel Times \parallel Nonce_3 \parallel Ticket_{rtgs} \parallel Authentica\ tor_c$

Where, $Authenticator_c = E_{K_{c,rtgs}}(ID_c \parallel \mathrm{Re}\,alm_c \parallel TS_2)$, ID_{rw} denotes identity of Remote web (*rw* or *RW*).

Message (8): Remote *TGS* returns ticket-granting ticket for Remote web (corresponding to (8) in Fig. 1):

$$RTGS \rightarrow C: \mathrm{Re}\,alm_c \parallel ID_c \parallel Ticket_{rw} \parallel E_{K_{c,rtgs}}(K_{c,rw} \parallel Times \parallel Nonce_3 \parallel \mathrm{Re}\,alm_{rw} \parallel ID_{rw})$$

Where, $Ticket_{rw} = E_{K_{rw}}(Flags \parallel K_{c,rw} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)$

Message (9): Client requests Remote web for remote service (corresponding to (9) in Fig. 1): $C \rightarrow RW : Options \parallel Ticket_{rw} \parallel Authentica\ tor_c$

Where, $Ticket_{rw} = E_{K_{rw}}(Flags \parallel K_{c,rw} \parallel \mathrm{Re}\,alm_c \parallel ID_c \parallel AD_c \parallel Times)$

$$Authentica\ tor_c = E_{K_{c,rw}}(ID_c \parallel \mathrm{Re}\,alm_c \parallel TS_3 \parallel Subkey \parallel Seq\#)$$

2) Justification analysis

Message (1) and Message (2) are the same as ones in the Local authentication.

In message (3) ID_{tgs} becomes ID_{rtgs}, namely *C* requests ticket-granting ticket for *RTGS*.

In message (4) $Ticket_{tgs}$ becomes $Ticket_{rtgs}$, namely *TGS* returns ticket-granting ticket of *RTGS*. $Reakm_{lw}$ and ID_{lw} become $Reakm_{rtgs}$ and ID_{rtgs}.

Message (7): When *C* presents this ticket $Ticket_{rtgs}$ to *RTGS*. *RTGS* can decrypt $Ticket_{rtgs}$ with K_{rtgs}: $D_{K_{rtgs}}(Ticket_{rtgs}) = D_{K_{rtgs}}(E_{K_{rtgs}}(Flags \| K_{c,rtgs} \| \mathrm{Re}alm_c \| ID_c \| AD_c \| Times))$

$$= Flags \| K_{c,rtgs} \| \mathrm{Re}\,alm_c \| ID_c \| AD_c \| Times$$

RTGS uses $K_{c,rtgs}$ to decrypt $Authenticator_c$:

$$D_{K_{c,rtgs}}(Authenticator_c) = D_{K_{c,rtgs}}(E_{K_{c,rtgs}}(ID_c \| \mathrm{Re}alm_c \| TS_2)) = ID_c \| \mathrm{Re}\,alm_c \| TS_2$$

RTGS can then check the name and address from $Authenticator_c$ with that of the $Ticket_{rtgs}$. If all match, then *RTGS* is assured that the sender of the $Ticket_{rtgs}$ is indeed the $Ticket_{rtgs}$'s real owner.

Message (8): *C* can gain $Ticket_{rw}$ for visiting Remote web; it can also obtain $K_{c,rw}$ by decrypting as follow: $D_{K_{c,rtgs}}(E_{K_{c,rtgs}}(K_{c,rw} \| Times \| Nonce_3 \| \mathrm{Re}\,alm_{rw} \| ID_{rw}))$

$$= K_{c,rw} \| Times \| Nonce_3 \| \mathrm{Re}\,alm_{rw} \| ID_{rw}$$

Message (9): Remote web can decrypt $Ticket_{rw}$ with K_{rw}:

$$D_{K_{rw}}(Ticket_{rw}) = D_{K_{rw}}(E_{K_{rw}}(Flags \| K_{c,rw} \| \mathrm{Re}\,alm_c \| ID_c \| AD_c \| Times))$$

$$= Flags \| K_{c,rw} \| \mathrm{Re}\,alm_c \| ID_c \| AD_c \| Times$$

Remote web uses $K_{c,rw}$ to decrypt $Authenticator_c$:

$$D_{K_{c,rw}}(Authentica\ tor_c) = D_{K_{c,rw}}(E_{K_{c,rw}}(ID_c \| \mathrm{Re}\,alm_c \| TS_3 \| Subkey \| Seq\#))$$

$$= ID_c \| \mathrm{Re}\,alm_c \| TS_3 \| Subkey \| Seq\#$$

Local web can check the name and address from $Authenticator_c$ with that of the $Ticket_{rw}$. If all match, *C* achieves authentication of Remote web. Remote web can also gain *Subkey*, and uses it to protect succeeding application session.

4 Security Analysis

The security of distributed web assurances provided by authentication service based on Kerberos can be attributed to five factors:

- *No-impersonating*

A network eavesdropper can not obtain the necessary information to impersonate a user. More generally, this authentication method is strong enough that a potential opponent does not find it to be the weak link.

- *Reliable*

Because Kerberos is highly reliable and employ distributed server architecture, with one system able to back up another [8], hence, the system has highly availability of the supported services and access control.

- *Transparent*

The user is only required to enter a password, he or she is not aware that authentication is taking place.

- *Scalable*

The system is capable of supporting large numbers of clients and web, because of modular and distributed architecture.

- *Resisting attack*

Rijndael encryption algorithm uses secure S-boxes as nonlinear components [8, 9]; it supports on-the-fly subkey computation for encrypting. The operations used by Rijndael are among the easiest to defend against power and timing attacks. The use of masking techniques to provide Rijndael with some defense against these attacks does not cause significant performance degradation. Because we adopt Rijndael as encryption function, our method has highly resisting attack.

5 Conclusions

ID authentication mechanism clears the way for increasing distributed web security and provides a powerful guarantee, which effectively prevents farther harm for computer network systems. In this paper, an ID authentication mechanism was presented, and this mechanism takes on no-impersonating, highly availability, transparent and scalable. Thus it has an extensive worthiness of applications and theories in network security field.

References

1. Seixas, N., Fonseca, J., Vieira, M.: Looking at Web Security Vulnerabilities from the Programming Language Perspective: A Field Study. Software Reliability Engineering 1, 129–135 (2009)
2. Vieira, M., Antunes, N., Madeira, H.: Using web security scanners to detect vulnerabilities in web services. In: IEEE/IFIP International Conference on dependable systems & networks, vol. 1, pp. 566–571 (2009)
3. Rehbock, S., Hunt, R.: Trustworthy clients: Extending TNC to web-based environments. Computer Communications 32(5), 1006–1013 (2009)
4. Basso, A., Sicco, S.: Preventing massive automated access to web resources Computers & Security, vol. 28(3-4), pp. 174–188 (2009)
5. Yamany, H.F.E.L., Capretz, M.A.M., Allison, D.S.: Intelligent security and access control framework for service-oriented architecture. Information and Software Technology 52(2), 220–236 (2010)
6. Han, S., Dillon, T., Chang, E.: Secure web services using two-way authentication and three-party key establishment for service delivery. Journal of Systems Architecture 55(4), 233–242 (2009)
7. Steiner, J.G., Neuman, C., Schiller, J.I.: Kerberos: An Authentication Service for Open Network Systems. In: Proceedings of the 1988 Winter USENIX Conference, pp. 191–202 (February 1988)
8. Whitman, M.E., Mattord, H.J.: Principles of Information Security, 3rd edn. Thomson Course Technology (2006)
9. Muda, Z., Mahmod, R., Sulong, M.R.: Key transformation approach for Rijndael security, pp. 290-297 (2010)

A New Scheme for Protecting Master-Key of Data Centre Web Server in Online Banking

Cao Lai-Cheng[*] and Liang Lei

School of Computer and Communication,
Lanzhou University of Technology, Lanzhou 730050, China
caolch@lut.cn

Abstract. The master-key is used to encrypt the operation-key, and the operation-key is applied to encrypt the transport-key, consequently safety protection of the master-key is security core in online banking system. A scheme to protect the master-key was presented. Using method of 3-out-4 key share and LaGrange formula, the shares of the master-key were distributed to one synthesizing card and four key servers. When the data centre web server needed the master-key, the synthesizing card firstly authenticated the legitimacy of the shares of randomly selected three key severs from the four by zero-knowledge proof technology, once the shares were modified and destroyed, rest shares could make up a group so that the system worked continuously. Then the synthesizing card synthesized the master-key based on the shares of those three key severs. Security analysis proves that this scheme makes the whole system to have fault-tolerant and error detection, and also shows no-information leakage and defending collusive attack.

Keywords: web security; master-key; online banking; fault-tolerant; zero-knowledge proof.

1 Introduction

Nowadays, more and more men fulfill electronic payment through the online banking, and then it becomes more and more important that the capital information is assured to be secure and not modified illegally. Security of user's confidential information lies on security of the encrypting key in the entire information exchange, and the key management mechanism of the online banking adopts three levels key management model.

- *Master-Key*: it is used to encrypt operation-key, its security level is highest. According to different demands of hardware encryption equipment, this key is generated from hardware encryption machine.
- *Operation-Key*: it is generated from the data centre web server (DCWS) of the online banking and distributed to the branch bank; its function is to encrypt

[*] This work is supported by the National Natural Science Foundation of China under Grant No. 60972078; the Gansu Provincial Natural Science Foundation of China under Grant No. 0916RJZA015.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 179–186, 2010.

different transaction data, such as, user's bank account number and password. This key is encrypted by master-key and stored in the DCWS with cryptograph format.

- *Transport-Key*: it is used to protect the transaction data between the DCWS and the branch bank or between the branch bank and the branch bank. This key, which is encrypted by the operation-key, can be provided from one party to the other party, or it is produced by mutual consultation.

From the above, we can see that the security of the master-key is a key and core of security of entire system, paper [1-3] point out the importance of the security of the master-key, and the online banking is subject to attacks at various resources, paper [4] introduces a command injection attack, which poses a serious threat to master-key. Chonka Ashley [5] presents DDOS attacks by SOTA to the DCWS. If attacker succeeds to crack the master-key, he can connect and totally control the DCWS from remote place. In this paper, we put forward scheme to protect the master-key of the DCWS in online banking based on fault-tolerant, this master-key is protected by distributing its shares to four key servers and one synthesizing card, and key servers whose shares are modified can be found by zero-knowledge proof technology.

The remainder of this work is organized as follows. In section II we describe our security architecture design. In Section III we present a scheme to protect the master-key of the DCWS in online banking by fault-tolerant. In section IV we finish security analysis about our method. The conclusion is presented in section V.

2 Security Architecture Design

The current online banking includes the DCWS, the database, the first class net, branch bank common gateway, tandem net, branch bank client, ATM client and posp client, its network topology is shown as in Fig. 1. The authentication process of electronic transaction of user is shown as following:

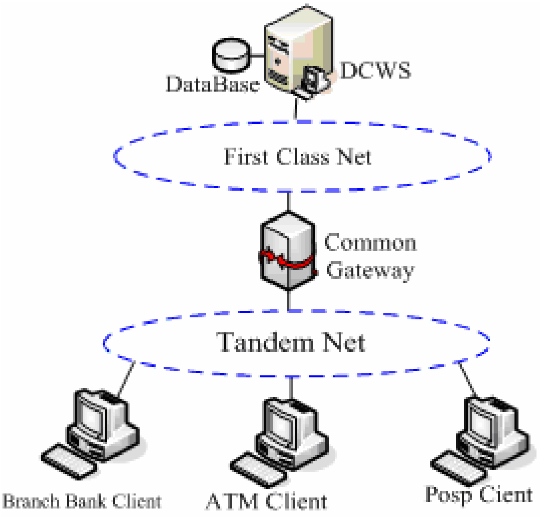


Fig. 1. Network topology of the current online banking

1) Encrypting user's information

User-card-number and user-password are encrypted by the operation-key of branch bank terminal with DES encryption algorithm, that is

$$CRY = Ecrypt - DES_{Operation-Key}(\text{User - Card - Number} \parallel \text{User - Password})$$

Then, branch bank terminal generates 128 bit message authentication code (*MAC*) of the cryptograph *CRY* with *MD5* algorithm, that is

$$MAC = MD5(CRY)$$

Finally, branch bank terminal sends *CRY* and *MAC* to the DCWS.

2) Authenticating integrality

The DCWS uses same MD5 algorithm to authenticate integrality of the message CRY.

Then, the DCWS decrypts the relevant cryptograph of the database with its master-key to acquire the operation-key of branch bank terminal.

Finally, the DCWS decrypts *CRY* with this operation-key to get the user-card-number and user-password, and contrasts with datum of the database, then finishes authentication of user.

But the master-key of the DCWS is protected by security strategy of the bank, because the security of the master-key is a key and core of security of entire system from the analysis in section I, it is undependable and insecure. Thus, we have extended this network topology of the online banking, a synthesizing card is embedded into the DCWS, and the DCWS connects four key servers with the hub, this master-key is protected by distributing its shares to these key servers and this synthesizing card, the security architecture is shown as in Fig. 2.

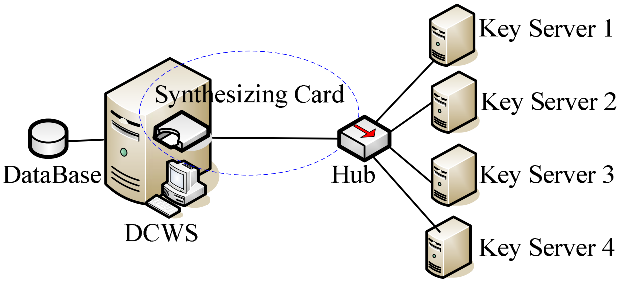


Fig. 2. The security architecture of the current online banking

3 The Scheme Protecting Master-Key

3.1 The Method of 3-Out-4 Key Share

In the method of t-out-n key share [6-8], the key is divided into t shares, which are stored in n share servers, only the key can be synthesized by t shares, this paper put forward arithmetic of synthesizing master-key based on fault-tolerant: 3 shares are stored in 4 key severs, and the shares are designed for 2 groups to increase fault tolerance. Table 1 shows this method of 3-out-4 master-key share, each key severs has two group shares. When any one key sever loses its shares, the system can continuously work. When any two key severs is attacked, any information of the master-key can not be leaked.

Table 1. The method of 3-out-4 master-key share

Group number	Key Sever 1	Key Sever 2	Key Sever 3	Key Sever 4
Group 1	d_{11}	d_{12}	d_{13}	d_{13}
Group 2	d_{21}	d_{21}	d_{22}	d_{23}
Compounding 1: d= d_{11}+ d_{12}+ d_{13}				
Compounding 2: d= d_{21}+ d_{22}+ d_{23}				

3.2 The Method of Distributing Shares

Referring to [4-5], the method of distributing shares is presented the follow.

1) Choosing a polynomial:

$$f(x)=\sum_{k=0}^{3} a_k x^k \tag{1}$$

2) Generating 4 random numbers x_i (i=0,1,...,3), and computing:

$$f(x_i)=\sum_{k=0}^{3} a_k x_i^k$$

3) When i=0, taking $d_0=f(x_0)$ as the share of synthesizing card.

4) When i=1, 2, 3, using LaGrange formula to computer:

$$f(x)=\sum_{i=1}^{3}(f(x_i)\sum_{j=1,3}^{j\neq i}\frac{x-x_j}{x_i-x_j}) \tag{2}$$

Let: $x=0$

$$f(0)=\sum_{i=1}^{3}(f(x_i)\sum_{j=1,3}^{j\neq i}\frac{-x_j}{x_i-x_j}) \tag{3}$$

5) Taking m (m=1, 2) group shares of each key sever to compute:

$$d_{mi}=f(x_i)\sum_{j=1,3}^{j\neq i}\frac{-x_j}{x_i-x_j} \qquad (i=1,2,3) \tag{4}$$

6) Then share datum of synthesizing master-key is:

$$d=d_0+\sum_{i=1}^{3}d_{mi} \tag{5}$$

3.3 Authenticating Legitimacy

Randomly selecting 3 key severs, the synthesizing card can authenticate legitimacy of each group shares by zero-knowledge proof technology before the master-key is synthesized, this performance can judge which group shares are modified and destroyed, the follow is authenticating process:

In advance, the public keys of each group shares in the key sever are stored into the database of the DCWS, and the arithmetic of the generated public key is: choosing two big prime number p and q, the public key H_{mi} of m (*m=1, 2*) group shares in the key sever is computed:

$$H_{mi} = p^{d_{mi}} \bmod q \quad (m=1, 2; i=1, 2, 3) \tag{6}$$

Table 2 shows the characteristic of the public keys of the key severs.

Table 2. The characteristic of the public keys

Characteristic	Key Sever 1	Key Sever 2	Key Sever 3	Key Sever 4
Group 1				
Public key	H_{11}	H_{12}	H_{13}	H_{13}
Share number	11	12	13	13
Group 2				
Public key	H_{21}	H_{21}	H_{22}	H_{23}
Share number	21	21	22	23

The synthesizing card selects an integer $x_k \in [1, p-1]$ and computes:

$$y_k = p^{x_k} \bmod q \tag{7}$$

Then it broadcasts y_k to the key severs.

The key severs compute:

$$p_{mi} = {y_k}^{d_{mi}} \bmod q \, (m=1, 2; i=1, 2, 3) \tag{8}$$

Then it sends p_{mi} to synthesizing server.

Using public key H_{mi} in the database, synthesizing card authenticates the follow equation:

$$p_{mi} = H_{mi}^{x_k} \bmod q \; (m=1, 2) \tag{9}$$

If the left value and right value of this equation are not equivalent, it is authenticated that the m (m=1, 2) group shares are modified or destroyed.

3.4 Synthesizing Master-Key

Case 1: If the synthesizing card has authenticated that 3 key severs are all legitimacy, it selects one group shares whose share numbers are sequential after they are arrayed again for share datum of synthesizing master-key.

Case 2: If the synthesizing card has authenticated that some shares are not legitimacy and the rest shares can form one group, when these share numbers are sequential after they are arrayed again, the synthesizing card selects this group for share datum of synthesizing master-key.

Case 3: If all legitimacy shares can not form one group, whose share numbers are sequential after they are arrayed again, then the synthesizing card selects another one group key severs.

Supposing the synthesizing card to select one group share numbers is m1, m2, m3 (m=1, 2), and then the arithmetic of synthesizing master-key is showed the follow.

First, the synthesizing card computes:

$$p_0 = y_k^{d_0} \bmod q \tag{10}$$

Second, the synthesizing card synthesizes the master-key:

$$\text{Master - Key} = p_0 \prod_{i=1}^{3} p_{mi} \tag{11}$$

Fig. 3 shows that the synthesizing card synthesizes the master-key to base on key sever 1, 2, 4.

Where, the synthesizing card using first group shares to synthesize the master-key is

$$\text{Master - Key} = p_0 \prod_{i=1}^{3} p_{1i}$$

The synthesizing card using second group shares to synthesize the master-key is

$$\text{Master - Key} = p_0 \prod_{i=1}^{3} p_{2i}$$

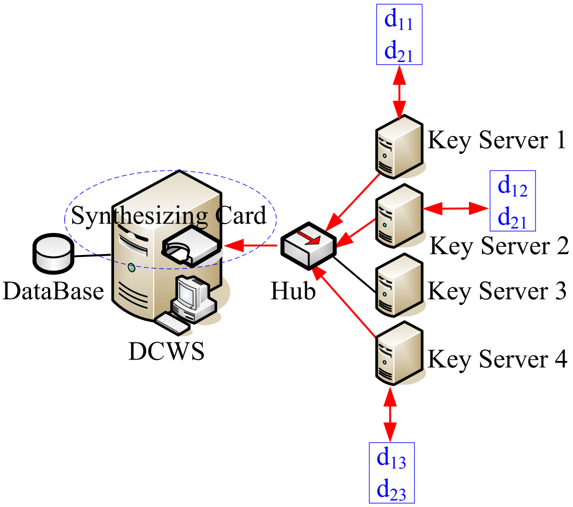


Fig. 3. Synthesizing master-key by key sever 1, 2, 4

- *Correctness proof*:

Let d_{mi} and d_{li} (m, l=1, 2; i=1, 2, 3) as two group shares, they synthesize the master-key K_m and K_l.

Because $d = d_0 + \sum_{i=1}^{3} d_{mi} = d_{0+} \sum_{i=1}^{3} d_{li}$, thus

$$K_m = p_0 \prod_{i=1}^{3} p_{mi} = y_k^{d_0} \bmod q \prod_{i=1}^{3} y_k^{d_{mi}} \bmod q$$

$$= y_k^{d_0 + \sum_{i=1}^{3} d_{mi}} \bmod q = y_k^d \bmod q$$

$$K_l = p_0 \prod_{i=1}^{3} p_{li} = y_k^{d_0} \bmod q \prod_{i=1}^{3} y_k^{d_{li}} \bmod q = y_k^{d_0 + \sum_{i=1}^{3} d_{li}} \bmod q$$

$$= y_k^d \bmod q$$

Therefore, $K_m = K_l$.

3.5 The Share Updating

The synthesizing card updates those shares in each period, namely the shares are distributed again in next period to rely on the method of section 3.2.

4 Security Analysis

- *Fault-tolerant*

According to equation (8) and (11), when one key sever loses its shares, the system can continuously work.

- *Error detection*

According to equation (6), (7), (8) and (9), if some shares of selected 3 key severs lose their shares or their shares are modified, the synthesizing card can all find these errors to base on zero-knowledge proof technology.

- *No-information leakage*

On the one hand, because equation (7) and (8) base on the computational infeasibility of discrete logarithms, only $y_k = p^{x_k} \bmod q$ and $p_{mi} = y_k^{d_{mi}} \bmod q$ can reflect the information of the master-key, it is impossible that using x_k and p_{mi} compute d_{mi}, namely, attackers can not obtain any share in key servers from filched p_{mi}.

Suppose an adversary tries to acquire as many of key shares as possible, he can not synthesize the master-key in next period because synthesizing card updates those shares in each period.

On the other hand, in equation (5), the share d_{mi} (m=1, 2; i=1, 2, 3) is leaked, it can not result in leaking share datum d. Because d_{mi} is randomly selected, d_{mi} and d are irrelative, namely, conditional information entropy $H(d \mid d_{mi}) = H(d)$, similarly,

d_{mi} and d_{mj} ($i \neq j; i, j = 1,2,3$) are unattached and random, then conditional information entropy $H(d \mid d_{mi}, d_{mj}) = H(d)$, thus the leak of a few d_{mi} (i=1,2,3) can not leak share datum d.

- *Defending collusive attack*

If three ones of four key severs perform collusive attack, they can select one group shares whose share numbers are sequential after they are arrayed again, but d_0 and d_{mi} ($i = 1,2,3$) are random and irrelative, and conditional information entropy $H(d_0 \mid d_{mi}) = H(d_0)$, in addition, participating in synthesizing the master-key x_k is randomly selected in [1, p-1], therefore, collusive attack can not finish.

5 Conclusions

This paper put forwards a scheme to protect the master-key of the DCWS in online banking system. We have demonstrated that scheme is robust and has four major contributions: ① Fault-tolerant. ②Error detection. ③ No-information leakage. ④ Defending collusive attack. This scheme can provide safety protection of the master-key of the DCWS, thus it has an extensive worthiness of applications and theories in security field of online banking system.

References

1. Guo, H., Mu, Y., Zhang, X.Y.: Enhanced McCullagh-Barreto identity-based key exchange protocols with master key forward security. International Journal of Security and Networks 5(2-3), 173–187 (2010)
2. Hua, G., Yi, M., Xiyong, Z.: Novel and efficient identity-based authenticated key agreement protocols from weil pairings. In: Zhang, D., Portmann, M., Tan, A.-H., Indulska, J. (eds.) UIC 2009. LNCS, vol. 5585, pp. 310–324. Springer, Heidelberg (2009)
3. Morrissey, P., Smart, N.P., Warinschi, B.: The TLS handshake protocol: A modular analysis. Journal of Cryptology 23(2), 187–223 (2010)
4. Zhendong, S., Gary, W.: The essence of command injection attacks in web applications. ACM SIGPLAN Notices 41(1), 372–382 (2006)
5. Ashley, C., Wanlei, Z., Yang, X.: Protecting web services from DDOS attacks by SOTA. In: ICITA 2008, pp. 379–384 (2008)
6. Wu, T., Malkin, M., Boneh, D.: Building intrusion-tolerant applications. In: Information Survivability Conference and Exposition, pp. 25–27. IEEE Computer Society, Los Alamitos (2000)
7. Xian-feng, Z., Jin-de, L.: A threshold ECC Based on Intrusion Tolerance TTP Scheme. Computer applications 24(2), 5–8 (2004)
8. Shoup, V.: Practical threshold signatures. In: Preneel, B. (ed.) EUROCRYPT 2000. LNCS, vol. 1807, pp. 207–220. Springer, Heidelberg (2000)

An Automated Worm Containment Scheme

Lipeng Song[1] and Zhen Jin[2]

[1] School of Electronic and Computer Science and Technology,
North University of China, 030051 Taiyuan, China
slp880@gmail.com
[2] School of Science, North University of China, 030051 Taiyuan, China
jinzhen@nuc.edu.cn

Abstract. How to detect and alleviate intelligent worms with the characteristic of both slow scanning rate and high vulnerability density? Here, we present a scheme to solve the problem. Different from previous schemes, which set a limit on instantaneous scanning rate against each host, the scheme considered in this paper counts the number of unique IP addresses contacted by all hosts of a subnet over a period and sets a threshold to determine whether the subnet is suspicious. Specially, we consider the similarity of information required by users belonging to the same subnet. The result shows that our scheme is effective against slow scanning worms and worms with high vulnerability density.

Keywords: network security, automatic worm containment, slow scanning worms.

1 Introduction

The internet has become more and more important to global economy and to people`s life. Meanwhile, computer worms have caused many serious problems to the security of internet. For example, the SQL Slammer worm infected over 90% of the vulnerable hosts on the Internet within ten minutes[1]. The Code Red II, which was released in 2001, infected more than 359000 machines during a period of less than 14 hours[2]. The cost of the epidemic is $2.6 billion estimated by Computer Economics. Thus, we need automated detection and timely response to defend against worms.

The rate throttling methods which set a limit on instantaneous scanning rate against all hosts are proposed in [3,4]. They are effective on fast scanning worms. However, it is generally accepted that they are not effective against slow scanning worms because it is especially difficult to set the limit without affecting the normal traffic [5]. The reason is that the difference of scanning rate between slow scanning worms and normal hosts is very tiny.

Sellke et al. [5] give a scheme which can prevent the spread of slow-scanning worms by limiting the number of distinct IP addresses contacted per host over a period. However, the scheme will be invalid when the density of vulnerable hosts is high.

The aim of this research is to give a scheme to constrain the spread of intelligent internet worms, which exploit both slow scanning rate and high vulnerability density.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 187–193, 2010.

The scheme is concerned with worms which infect hosts by scanning a list of randomly generated IP addresses. Those worms such as e-mail worms and peer-to-peer worms which use other methods to find vulnerable hosts are not included in this scheme.

Ma et al. [6] analyze the characteristic of user information requirements in a region. They discover that the web sites visited by the users in one region follow the power law and have aggregated phenomenon.

In this paper, we present a worm containment scheme which counts the number of distinct IP addresses contacted by a subnet over a period and set a value as the threshold. The subnet whose total contacting number exceeds the threshold is simply considered to be suspicious and then its scanning rate will be limited to a given rate. The result shows that we can easily give the threshold which is effective in detecting the intelligent worms but has marginal influence on the normal traffic. The reason is that the normal hosts in one subnet show similarity in visiting web sites but infected hosts do not and thus the aggregated difference of distinct IP addresses contacted by clean subnet and by infected subnet becomes very large.

The main contributions of this paper are summarized as follows: We present a scheme to automatically constrain the spread of worms with both slow scanning rate and high vulnerability density. We consider the aggregated phenomenon of user visiting web sites and apply it to worm detection system for the first time.

The remainder of this paper is organized as follows: Section 2 reviews relevant research on network worms. Section 3 gives the worm containment scheme and numerical results. Section 4 analyzes the impact of dynamic immunization. We summarize in section 5.

2 Related Work

One basic model used to study worm propagation, which assumes that there is no patching and the infection rate is constant, is the Random Constant Spread (RCS) model proposed by Staniford et al.[7]:

$$dI(t)/dt = \beta I(t)(N - I(t)), \tag{1}$$

where N, I(t), β represent the total number of susceptible hosts on the Internet, the total number of infected hosts at time t, and the constant infection rate, respectively.

As mentioned in Section 1, rate control schemes have modified the RCS model by adding a limiting factor on instantaneous scanning rate against all hosts [3,4]:

$$dI(t)/dt = \beta_1 I(t)(N - I(t)), \tag{2}$$

where β_1 is the infection rate allowed by limiting factor and $\beta >> \beta_1$. These schemes are effective in slowing down fast worms. However, they are not effective against slow scanning worms.

Worm detection systems have been proposed by several researchers. Zou et al.[8] use a Kalman filter to detect the worms, which can separate worm traffic from background nonworm scan traffic. Kabiri et al.[9] develop a worm detection system called NIDS in which detection knowledge is maintained manually to detect the worms. To automatically acquire detection knowledge, machine Learning techniques are used in [10,11].

Fu et al. [12] give a model considering the effect of neighbor-alarm. The model requires that an immune node must send an alarm to its neighbors. However, the model will lose its suppression function if its neighbors do not send alarms.

3 Worm Containment Scheme

In this section, we first present the worm containment scheme and then give some simulation results. The worm containment system is based on the idea we have discussed in section 1. Let M_s be the threshold, total number of distinct IP addresses allowed for a subnet over a given period, and S_f represent the given rate allowed for a suspicious subnet. The worm containment scheme is given bellow:

I. For each subnet, set a counter to count the number of distinct IP addresses initiated by them. At the beginning, each counter is set to zero.

II. Increment the counter for a subnet when the subnet initiates a new IP scanning.

III. If a subnet reaches the threshold, it`s scanning rate will be limited to S_f for the rest of the period.

IV. At the end of the period, release the rate control for those suspicious subnets. Reset each counter to zero and then return to I.

Note that the values of M_s and S_f have great influence on the propagation of worms. Then, we can predefine the allowed infection percentage to determine these values, which will be seen in our simulation.

According to the trace data from NLANR [13], there are 46k distinct IP addresses contacted by 400 people in Bell Labs during one week; that is, the total number of distinct IP addresses contacted by the subnet per hour is 280. However, the total number of IP addresses contacted those people is over 60k.

Although our scheme is effective with various subnets, we suppose, for simplicity, that each subnet has 400 end hosts and there is only one public IP address assigned to a web proxy. When a web proxy is infected by a worm, with a high probability (50% in our simulation) we suppose the inner hosts are also susceptible to the worm. Moreover, we omit the inner propagation time since it is much faster.

We also suppose that the worm`s scanning rate is 2 per hour which is below the scanning rate of top 10 people in [13]. Let P=0.005 represent the vulnerability density. Then, the infection rate denoted as β is 0.01 per hour.

We first simulated the propagation of a random scanning worm with our containment system. Fig. 1 gives the simulation results with three different thresholds (M_s).

Fig. 1 shows that our containment system is effective in constraining the slow scanning and high vulnerability worms. It yields a 2/3 slowdown compared with the RCS system.

Fig. 1 also shows there is very little difference when we vary the threshold. The reason is that all of the thresholds are small enough for limiting the worms`s propagation.

To see the influence of the allowed scanning rate for suspicious subnet, we simulated our containment system with various S_f. The results are given in Fig. 2.

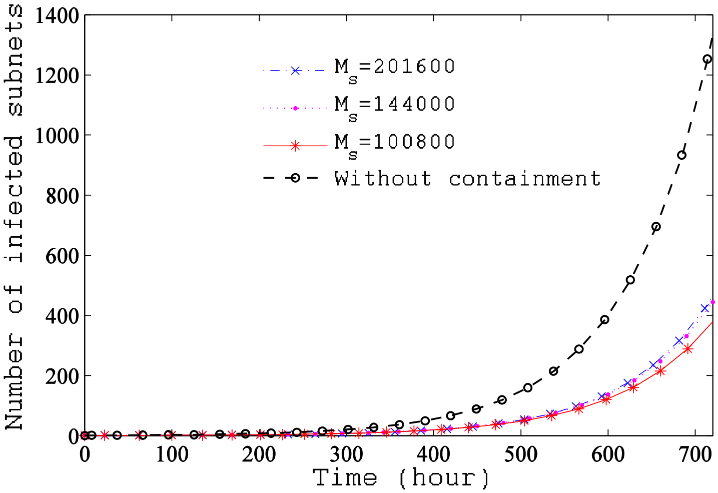


Fig. 1. Number of infected subnets during 30 days with one initially infected subnet. S_f is set to 280 per hour the same as normal subnet's contacting rate.

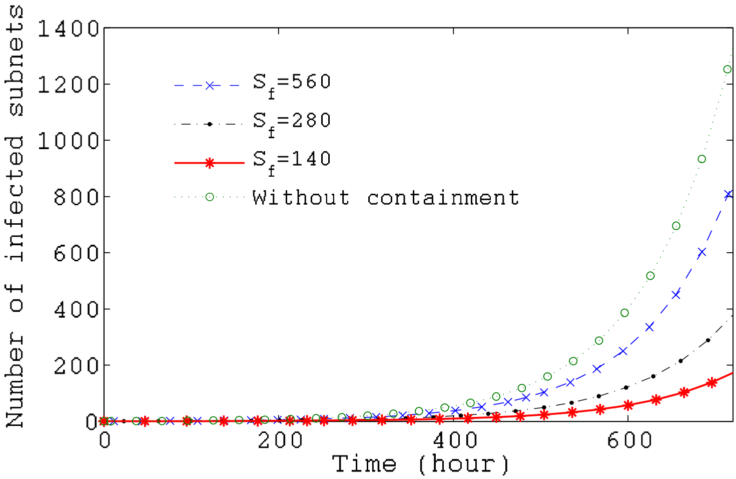


Fig. 2. Numbers of infected subnets during 30 days with one initially infected subnet and M_s=100800.

As shown in Fig. 2, the total number of infected subnets is less than 200 when S_f=140, which is only 1/7 of the RCS system`s. Fig. 2 also shows that the total number of infected subnets will increase proportional to the increase of $S_f$. However, the throttling effect is still obvious even when $S_f$=560, two times of the normal subnet`s contacting rate.

4 The Effect of Dynamic Immunization

In previous section we do not consider the effect of patching and dynamic immunization. However, anti-virus program will dynamically scan the vulnerabilities exploited by worms. The infection progress will be impeded if the vulnerabilities are patched.

To model the effect of dynamic patching on the propagation of worms, we introduce a geometric time-to-repair (TTR) policy [14] in our containment system. With independent probability δ_R on each time step, an infected subnet is returned to the recovered state and immunized against infection.

We simulated our containment system with δ_R =0.003, which means the mean time to repair is two weeks. Fig. 3 gives the simulation results.

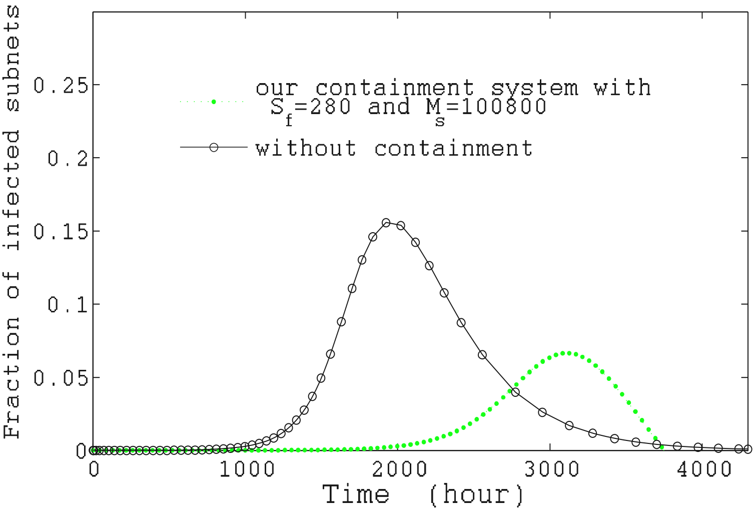


Fig. 3. The fraction of infected subnets of our containment system is compared with the RCS system's

As our simulation results show, there is over 15% susceptible subnets infected in the RCS system when the dynamic immunization is considered. Our containment system is effective on constraining the worm`s propagation. It has a 50% slowdown compared with the RCS system. Furthermore, our system delays the coming of the peak point, which is of benefit to the anti-virus efforts.

We also simulated our system with various δ_R to see the influence of repair rate on the propagation of worms. Fig. 4 shows the simulation results with M_s =100800 and S_f =280.

Fig. 4 shows that the infection rate of peak point is over 20% when the mean repair time is a month (δ_R=0.0014). It decreases to 7% if the mean repair time is two weeks (δ_R=0.003), which demonstrates the great influence of the repairing rate on the worm`s propagation.

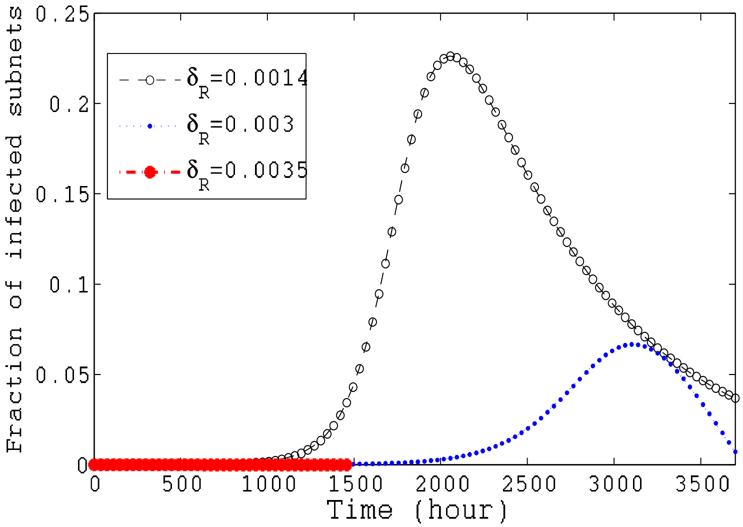


Fig. 4. The fraction of infected subnets of our containment system with δ_R =0.0014, 0.003 and 0.0035, the corresponding repair time is a month, two weeks and twelve days, respectively

From Fig. 4, we also find that the random scanning worm will be terminated at the beginning of the simulation if the mean repair time is less than two weeks. It is very important for us to control the propagation of random scanning worms.

5 Conclusion

Recently, the researches concerning network security and malware have focused on the games between antivirus system and virus, such as timely intrusion detection [15] and automated containment [5]. However, they are not effective on the slow scanning and high vulnerability density worms.

In this paper, we have exploited the similarity of information requirement among the users of the same subnet to provide the best method to detect and constrain the slow scanning and high vulnerability density worms. The similarity of information requirement leads to the aggregated of IP address contacted by those users, while the randomness characteristic of scanning worms will not. Based on this phenomenon, we design a system to automatically discriminate the infected subnets from normal subnets and then constrain the infected subnets.

Our results have shown that our containment system is effective on throttling the propagation of the slow scanning and high vulnerability density worms whether the dynamic immunization is adopted or not. More specifically, the threshold used to detect the suspicious subnets can be easily given since the difference of aggregated distinct IP addresses between normal subnet and infected subnet is very large. We have also obtained the influence of M_s, S_f and δ_R on the propagation of random scanning worms, which provides valuable information for antivirus system. The method presented can be easily adapted to constrain the propagation of mobile computer malware [16] and phone viruses [17].

In the future, we would like to provide a statistical model for the information requirement correlation among users of the same subnet. We would also like to evaluate our containment system using real data from enterprise networks. Our study was limited to the propagation of random scanning worms. We shall study the propagation model of topology-aware worms and then provide containment system for such worms.

References

1. Moore, D., Paxson, V., Savage, S., Shanon, C., Staniford, S., Weaver, N.: Inside the slammer wor. IEEE Security and Privacy journal (2003)
2. Moore, D., Shanon, C.: The Spread of the Code-Red Worm(CRv2) (2001), `http://www.caida.org/research/security/code-red/#crv2`
3. Williamson, M.M.: Throttling viruses: Restricting propagation to defeat malicious mobile code. Technical Report HPL-2002-172, HP Laboratories Bristol (2002)
4. Wong, C., Wang, C., Song, D., Bielski, S., Ganger, G.R.: Dynamic Quarantine of Internet Worms. In: Proc. IEEE Int'l Conf. Dependable Systems and Networks, pp. 73–82 (2004)
5. Sarah, S.H., Shroff, N.B., Bagchi, S.: Modeling and Automated Containment of Worms. IEEE Transcations on Dependable and Secure Computing 5, 71–86 (2008)
6. Ma, W.D., Wang, L., Li, Y.P., Shui, H.S., Zhou, M.T.: Influence of user requirement behaviors on internet collective dynamics. Acta Phys. Sin. 57, 1381–1388 (2008)
7. Staniford, S., Paxson, V., Weaver, N.: How to Own the Internet in Your Spare Time. In: Proc. Usenix Security Symp., pp. 149–167 (2002)
8. Zou, C.C., Gong, W., Towsley, D.: Monitoring and Early Warning for Internet Worms. In: Proc. ACM Conf. Computer and Comm. Security, pp. 190–199 (2003)
9. Kabiri, P., Ghorbani, A.A.: Research on Intrusion Detection and Response: A Survey. International Journal of Network Security 1, 84–102 (2005)
10. Kolter, J.Z., Maloof, M.A.: Learing to detect malicious executables in the wild. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 470–478 (2004)
11. Moskovitch, R., Gus, I., Pluderman, S., Stopel, D., Feher, C., Glezer, C., Shahar, Y., Elovici, Y.: Detection of Unknown Computer Worms Activity Based on Computer Behavior using Data Mining. In: Proc. IEEE Symposium on Computational Intelligence and Data Mining, pp. 202–209 (2007)
12. Fu, J.M., Chen, B.L., Zhang, H.G.: A Worm Containment Model Based on Neighbor-Alarm. In: Xiao, B., Yang, L.T., Ma, J., Muller-Schloer, C., Hua, Y. (eds.) ATC 2007. LNCS, vol. 4610, pp. 449–457. Springer, Heidelberg (2007)
13. Nlanr.: Bell Lab-I Data Set (2007), `http://pma.nlanr.net/Traces/long/bell.html`
14. Debany Jr., W.H.: Modeling the Spread of Internet Worms Via Persistently Unpatched Hosts. IEEE Netw 22, 26–32 (2008)
15. Stephenson, B., Sikdar, B.: A Quasi-Species Model for the Propagation and Containment of Polymorphic Worms. IEEE Trans. Computers 58, 1289–1296 (2009)
16. Hypponen, M.: Malware Goes Mobile. Scientific American 295, 70–77 (2006)
17. Wang, P., Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding the Spreading Patterns of Mobile Phone viruses. Science 324, 1071–1076 (2009)

RBAC-Based Access Control Integration Framework for Legacy System

He Guo[1,2], Guoji Lu[1], Yuxin Wang[1], Han Li[1], and Xin Chen[2]

[1] School of Computer Science and Technology
[2] School of Software Dalian University of Technology
Dalian, 116024, China
guohe@dlut.edu.cn, lugj207@163.com, wyx@dlut.edu.cn,
lihan409@gmail.com, chenx_dlut@163.com

Abstract. Access control comprises different kinds of access control policies. This paper proposed an RBAC-based access control integration framework to achieve and manage various access control policies during legacy system integration. Permission is defined as tasks, and tasks are extracted and organized as tree structure for each system. Then, a global task tree and an integrated policy library are generated for the integrated system to reorganize access control policies of different legacy systems. Additionally rules for authorization management are given to carry out further authorization. A case study is demonstrated to depict the proposed framework is a feasible and flexible solution for access control integration.

Keywords: legacy system; access control; system integration; Role-Based Access Control (RBAC); task tree.

1 Introduction

Access control is a mechanism which enables an authority to control access to some resources in a computer-based information system. It often protects resources in a software system so that these resources can only be accessed by authorized users. As most organizations nowadays store their sensitive data in software systems, access control is applied as a requisite technology in nearly every system.

Legacy systems are large software systems that continue to be used because of the cost of replacing and redesigning [1]. Along with the actual requirements of organizations, a variety of legacy systems require to be integrated. Different legacy systems may make organize access control policies. For the purpose of reducing the usage and management cost of the integrated system, a unified access control needs to be constructed by integrating various access control policies. Since role-based access control (RBAC) reduces the administrative operations by 50% to 90% over other traditional access controls [2], it is considered as the basis of this paper.

Based on RBAC, an access control integration framework for legacy systems is proposed. Firstly, tasks are extracted from each legacy system and a global task tree is generated for the whole integrated system. Then the original access control policies of each legacy system are transformed into the unified access control structure. Finally, rules for access control management are used to achieve further authorizations.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 194–201, 2010.

2 Related Work

Compared with other traditional access controls, RBAC is considered as an effective way to solve the resource control of large enterprises [3]. However, RBAC is primarily used during the development of new software systems, and it may not be applied in legacy systems. Recent years, some research focuses on the investigation of integrating RBAC into software systems.

In 2007, an agent-based service oriented approach to enforce existing systems with RBAC is claimed in [4]. In [5], an improved RBAC model based file protection system is designed. In addition, RBAC is also used for the interaction with search engines in a web-based learning management system (LMS) [6].

Except for integrating RBAC into software system, the administration of RBAC is also claimed to be a vital issue during system integration [7]. A model and its corresponding administration procedure for administrating RBAC in a distributed system are produced in [8]. Besides, [9] proposed a generic administration model for efficiently administrating different access control models.

Based on the above research, an RBAC-based access control integration framework is studied for the purpose of providing a flexible and convenient access control solution when legacy systems are integrated.

3 RBAC-Based Access Control Integration Framework

Fig. 1 illustrates the proposed framework, which consists of 4 modules as follows:

- Legacy Systems contains all legacy systems to be integrated.
- Task Library is composed of tasks extracted from legacy systems, where a task refers to a system function which is expressed as an interactive component in a legacy system.
- Integrated Policy Library is made up of 7 major tables comprising system, user, administrator, permission, role, user-role and role-permission.

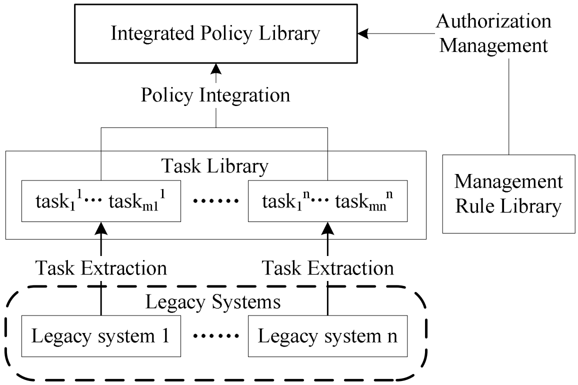


Fig. 1. Structure diagram of the proposed framework

- Management Rule Library involves the rules for further authorization.

In order to link the above modules, 3 operations are defined as follows:

- Task Extraction takes charge of extracting tasks from legacy systems using reverse engineering technology.
- Policy Integration is responsible for transforming different access control policies into tables defined in Integrated Policy Library.
- Authorization Management is in charge of performing corresponding authorizations according to the rules defined in Management Rule Library.

3.1 Task Extraction

Permission, which is defined as an action towards an object in access control, is always accomplished by a system function of a software system. Since almost all system functions of a software system are invoked by interactive components which refer to components that control the execution process of the system, permissions of access control are able to be reflected by interactive components. Thus, an interactive component whose corresponding system function implies permission is defined as a task, and the mapping from a user to a task is considered as a concrete access control policy. Table. 1 lists 4 tasks including Button1, Button2, Link1 and MenuItem1.

Table 1. Some examples of tasks

Object	Query	Update
SaleList.db	Button1	MenuItem1
Readme.txt	Link1	Button2

In order to extract tasks from a legacy system, static analysis is applied to the source code. Since task extraction depends on source code, the same tasks may be defined as different components in different languages. Take Button for example, it is implemented by CButton in C++ and JButton in java. Moreover, Hyperlink is regard as an interactive component if a legacy system is a web-based system.

For each legacy system, tasks are identified by traversing specific interactive components of the system. The result of task extraction is a global task tree, and the process of task extraction is depicted as follows.

Step 1. User Interface (UI) diagram generation: A UI class diagram is generated from the source code of each legacy system using Enterprise Architect (EA) which is a UML analysis and design tool.

Step 2. Task tree initialization: The name of the legacy system is set to be the root of the task tree, and the root is considered as the first level of the task tree.

Step 3. Task tree creation: Following steps are used to generate a task tree for a legacy system.

a) Second level generation: Interactive components contained in the user interface class C_{first} which is firstly visited by users are added to second level of the task tree, and C_{first} is appended to the current UI Class Set $C_{current}$.

b) Third level generation: For each UI class C_i in $C_{current}$, UI classes C_{set}^{i} which derive from C_i are found out according to the UI class diagram. For each UI class C in C_{set}^{i} the interactive component *ICom* which instantiate C is identified from C_i. Then interactive components defined in C are added as the descendants of *ICom*. Therefore, the third level of the task tree is built, and $C_{currentt}$ is replaced by the union of all C_{set}^{i}.

c) Subsequent levels generation: Repeat b) until $C_{current}$ is empty.

Step 4. Global task tree construction: The root of the global task tree is named as IS, and the global task tree is generated by combining the task trees of all legacy systems.

Based on the global task tree which contains all available tasks of all legacy systems, a specific task tree is generated for each user according to the permissions which are held by the user, and whether a user can access a task or not depends on the availability of the corresponding interactive component.

3.2 Policy Integration

Integrated Policy Library. In order to organize access control policies with RBAC, integrated policy library is constituted by 7 tables whose name and primary attributes are described in Table. 2. System table consists of the information of each legacy system. User table records the information of all users. Administrator table comprises the information of all who takes charge of managing integrated policy library. Permission table contains all tasks generated in task extraction. Role table covers all roles created after access control integration. User-role comprises the mapping from each user to all roles he holds. Role-permission stores the mapping from each role to his corresponding task tree. Moreover, for the purpose of distinguishing which legacy system a tuple belongs to, the ID of legacy system (lsid) is added as an attribute to user, role and permission table.

Table 2. Description of integrated policy library

Table name	Primary attributes
legacy system	(lsid, lsname)
user	(uid, lsid, uname)
administrator	(aid, lsid, aname)
role	(rid, lsid, rname)
Permission	(pid, lsid, pname)
user-role	(uid, rid)
role-permission	(rid, pid)

Sets and operations for access control integration. To achieve access control integration, sets (U, AU, R, P, LS, RU, RP) are defined to describe the essential information of tables contained in integrated policy library. For the purpose of manipulating the above sets, 6 operations are defined which is described in Table. 3, where a∈AU, perm∈P, role∈R.

Table 3. Definitions of operations

Operation	Description
CreateRole(a,role)	Create a role
DeleteRole(a,role)	Delete a role
AddPermission(a,perm,role)	Add a permission to a role,
RemovePermission(a,perm,role)	Remove a permission from a role
AddRole(a,user,role)	Add a role to a user
Remove(a,user,role)	Remove a role from a user

For a specific legacy system ls in LS, its corresponding U, R, P, UR, RP are respectively represented as ls.U, ls.AU, ls.R, ls.P, ls.UR and ls.RP. For instance, ls.U is the subset of U, which contains the user names of ls.

Transformation of access control policies. Transformation aims at reorganizing various kinds of access control using tables defined in integrated policy library. Before transforming access control policies, the original permissions of all legacy systems are represented by tasks. Then transformation is applied to each legacy system. Following discusses the transformation processes of the most common used types of access control.

a) Access control is achieved by a pre-defined username/password: A user is able to access the system if he/she holds the pre-defined username and password. A special case is that no username/password is pre-defined and the system can be accessed by any user. The formal description of the transformation is denoted as follows:

CreateRole (a, role),

$\forall$perm$\in$ls.P, AddPermission (a, perm, role);

$\forall$user$\in$ls.U, AddRole (a, user, role);

b) User groups are pre-defined to carry out access control: A user who can access the system belongs to at least one of the predefined user groups. The transformation is formalized as follows, where UG is a set of user groups, ug is a user group in UG, PUGug represents the mapping from ug to all its permissions and UUGug contains the mapping from ug to all its users.

$\forall$ug$\in$ls.UG, CreateRole (a, role);

$\forall$perm$\in$ls.PUGug, AddPermission (a, perm, role);

$\forall$user$\in$ls.UUGug, AddRole(a, user, role);

c) Access control is implemented by RBAC: Permissions related attributes in tables (permission and role-permission) are updated by tasks extracted from the source code.

3.3 Authorization Management

In order to achieve further authorization, 6 rules of management are given in this section, where predicate Enable(operation) is used indicate the operation is allowed to be executed.

Rule 1: The system administrator of a legacy system can create a role in the legacy system.

$a \in$ ls.AU => Enable (CreateRole (a, role)).

Rule 2: The system administrator of a legacy system can delete an exsiting role in the legacy system.

$a \in$ ls.AU $\wedge$ role $\in$ ls.R => Enable (DeleteRole (a, role)).

Rule 3: The system administrator of a legacy system can add permission to an exsiting role in the legacy system.

$a \in$ ls.AU $\wedge$ role $\in$ ls.R $\wedge$ perm $\in$ ls.P => Enable (AddPermission (a, perm, role)).

Rule 4: The system administrator of a legacy system can remove permission from an exsiting role in the legacy system.

$a \in$ ls.AU $\wedge$ role $\in$ ls.R $\wedge$ perm $\in$ ls.P $\wedge$(role, perm) $\in$ ls.RP => Enable (RemovePermission (a, perm, role)).

Rule 5: The system administrator of a legacy system can add a role to a user in the legacy system.

$a \in$ ls.AU $\wedge$ role $\in$ ls.R $\wedge$ user $\in$ ls.U =>Enable (AddRole (a, user, role)).

Rule 6: The system administrator of a legacy system can remove a role from a user in the legacy system.

$a \in$ ls.AU $\wedge$ role $\in$ ls.R $\wedge$ user $\in$ ls.U $\wedge$ (user, role) $\in$ ls.UR =>Enable (RemoveRole (a, user, role)).

4 Case Study

In order to ensure the feasibility of the proposed framework, An image and video processing system (abbr. IVPS) is chosen as the case study, which contains three user groups. Visual Studio 6.0 is used as the development environment.

Firstly, the UI class diagram of IVPS should be generated by EA. Based on the UI class diagram, task extraction is performed. During task tree initialization, IVPS is set to be the root and the first level of the task tree. As UI class FrmLogin is the interface which is first visited by users, it is considered as C_{first}, and the interactive components contained in FrmLogin is found out and forms the second level of the task tree. Within UI class FrmMain which is instantiated by the event of Button Login in FrmLogin, 4 interactive components are searched out and the third level of the task tree is constructed. After that, subsequent levels of the task tree are generated based during task extraction. Fig. 2 depicts the task tree of IVPS, where each node refers to a task and the max depth of tree is 5.

The access control of IVPS is accomplished based on three user groups (ug1, ug2 and ug3). 2 administrators (Andy and Bob) and 6 ordinary users (David, Jack, Mike, Paul, Roy and Tom) are selected in the case study. The original access control policies of IVPS are represented as follows, where childnodes(N) refers to all the descendants of node N.

IVPS.UG= {ug1, ug2, ug3};

IVPS.UUG=IVPS.UUG$_{ug1}$ $\cup$ IVPS.UUG$_{ug2}$ $\cup$ IVPS.UUG$_{ug3}$={(ug1,David,Jack, Mike), (ug2,Paul),(ug3,Roy,Tom)};

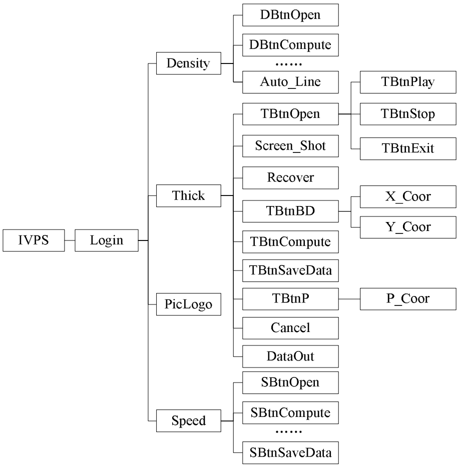


Fig. 2. Task tree of IVPS

IVPS.PUG={(ug1, Login, Density, PicLogo, childnodes(Density)), (ug2, Login, Thick, PicLogo, childnodes(Thick)), (ug3, Login, Speed, PicLogo, childnodes(Speed))};

According to the second access control transformation, r1, r2 and r3 are created. Following represents the transformed RBAC-based access control policies of IPVS.

IVPS.AU = {Andy, Bob}; IVPS.R= {r1, r2, r3};

IVPS.U = {David, Jack, Mike, Paul, Roy, Tom};

IVPS.P = {Login, Density, Thick, Speed, PicLogo …};

IVPS.UR = {(David,r1),(Jack,r1),(Mike,r1),(Paul r2),(Roy,r3),(Tom,r3)};

IVPS.RP = {(r1, Login, Density, PicLogo, childnodes(Density)), (r2, Login, Thick, PicLogo, childnodes(Thick)), (r3, Login, Speed, PicLogo, childnodes(Speed))}

Following example shows the process of generating a new role which can perform tasks X_Coor and Y_Coor, in which parentnodes(N) refers to all the predecessors of node N.

Firstly, r4 is created using Rule1 as follows:

Andy$\in$IVPS.AU => Enable (CreateRole (Andy, r4)).

Secondly, corresponding tasks are assigned to r4 as follows:

$\forall$perm$\in$X_Coor $\cup$ Y_Coor $\cup$ parentnodes(X_Coor) $\cup$ parentnodes(Y_Coor),Andy$\in$IVPS.AU$\wedge$r4$\in$IVPS.R =>Enable(AddPermission (Andy,perm, r4)).

When the above authorization completes, IVPS.R and IVPS.RP are updated as follows. Meanwhile relevant tables in integrated policy library are updated.

IVPS.R= {r1, r2, r3, r4};

IVPS.RP = {(r1, Login, Density, PicLogo, childnodes(Density)), (r2, Login, Thick, PicLogo, childnodes(Thick)), (r3, Login, Speed, PicLogo, childnodes(Speed)), (r4, Login, Thick, X_Coor, Y_Coor)};

In brief, the access control policies of the case study are organized by the integrated policy. All these reorganized access control policies satisfy the system requirements, and further authorizations can be applied on the basis of rules defined in management authorization.

5 Conclusion

An RBAC-based access control integration framework is proposed to organizing various access control policies in a unified structure in this paper. Firstly, permission of all legacy systems is replaced by tasks, and a global task tree is generated during task extraction. Secondly, integrated policy library is defined to preserve different access control policies in the same structure. Next, sets and operations for access control integration are defined to carry out access control policy transformation. Finally, 6 rules of management are proposed to accomplish further authorizations. The presented case study proves that the proposed access control integration framework for legacy systems is feasible, effective and flexible.

References

1. Bennett, K.H.: Legacy System: Coping with Success. IEEE Software 12(1), 19–23 (1995)
2. Antonio, A.: Migrating to Role Based Access Control (2006), `http://www.altametric.com/journal/EntCasePub.pdf`
3. Bertino, E.: RBAC models–concepts and trends. Computers & Security 22(6), 511–514 (2003)
4. Chen, F., Li, S., Yang, H.: Enforcing Role-Based Access Controls in Software Systems with an Agent Based Service Oriented Approach. In: Proceedings of the 2007 IEEE International Conference on Networking, Sensing and Control, London, pp. 15–17 (2007)
5. Ke, G., Ling, J., Hao, Y., Liao, H., Yang, Z.: Research and implementation of file protection system based on improved role-based access control. In: International Symposium on Computational Intelligence and Design, Changsha, pp. 242–245 (2009)
6. Bozzon, A., Iofciu, T., Nejdl, W., Taddeo, A.V., Tonnies, S.: Role Based Access Control for the interaction with Search Engines. In: Proceedings of the 1st International Workshop on Collaborative Open Environments for Project-Centered Learning, Greece, pp. 24–33 (2007)
7. Li, N., Mao, Z.: Administration in Role-Based Access Control. In: 2nd ACM Symposium on Information, Computer and Communications Security, Singapore, pp. 127–138 (2007)
8. Dekker, M.A.C., Crampton, J., Etalle, S.: RBAC Administration in Distributed Systems. In: Proceedings of ACM Symposium on Access Control Models and Technologies, Estes Park, pp. 93–101 (2008)
9. Li, X., Feng, D., Xu, Z.: A Generic Access Control Administration Model. Journal of Computer Research and Development 44(6), 947–957 (2007)

A Pseudo Random Numbers Generator Based on Chaotic Iterations: Application to Watermarking

Christophe Guyeux, Qianxue Wang, and Jacques M. Bahi

University of Franche-Comte, Computer Science Laboratory LIFC,
25030 Besançon Cedex, France
{christophe.guyeux,qianxue.wang,
jacques.bahi}@univ-fcomte.fr

Abstract. In this paper, a new chaotic pseudo-random number generator (PRNG) is proposed. It combines the well-known ISAAC and XORshift generators with chaotic iterations. This PRNG possesses important properties of topological chaos and can successfully pass NIST and TestU01 batteries of tests. This makes our generator suitable for information security applications like cryptography. As an illustrative example, an application in the field of watermarking is presented.

Keywords: Internet Security; Chaotic Sequences; Statistical Tests; Discrete Chaotic Iterations; Watermarking.

1 Introduction

The extremely fast development of the Internet brings growing attention to information security issues. Among these issues, the conception of pseudo-random number generators (PRNGs) plays an important role. Secure PRNGs which can be easily implemented with simple software routines are desired. Due to the finiteness of the set of machine numbers, the sequences generated by numerous existing PRNGs are not actually random. For example, the use of stringent batteries of tests allows us to determine whether these sequences are predictable. Chaos theory plays an active role in the improvement of the quality of PRNGs [5], [14]. The advantage of using chaos in this field lies in its disordered behavior and its unpredictability.

This paper extends the study initiated in [3] and [17]. In [3], it is proven that chaotic iterations (CIs), a suitable tool for fast computing iterative algorithms, satisfy the topological chaotic property, as it is defined by Devaney [7]. In [17], the chaotic behavior of CIs is exploited in order to obtain an unpredictable behavior for a new PRNG. This generator is based on chaotic iterations and depends on two other input sequences. These two sequences are generated by two logistic maps. Our generator has successfully passed the NIST (National Institute of Standards and Technology of the U.S. Government) battery of tests. However it appeared that it is a slow generator and it can't pass TestU01 because of the input logistic maps. Moreover this logistic map has revealed serious security lacks, which make it use inadequate for cryptographic applications [1].

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 202–211, 2010.

That is why, in this paper, we intend to develop a new fast PRNG. It will pass TestU01, widely considered as the most comprehensive and stringent battery of tests. This goal is achieved by using the ISAAC and XORshift maps in place of the two logistic maps. Chaotic properties, statistical tests and security analysis [19] allow us to consider that this generator has good pseudo-random characteristics and is capable to withstand attacks.

The rest of this paper is organized in the following way: in Section 2, some basic definitions concerning chaotic iterations and PRNGs are recalled. Then, the generator based on discrete chaotic iterations is presented in Section 3. Section 4 is devoted to its security analysis. In Section 5, we show that the proposed PRNG passes the TestU01 statistical tests. In Section 6 an application in the field of watermarking is proposed. The paper ends by a conclusion and some discussions about future work.

2 Basic Recalls

This section is devoted to basic notations and terminologies in the fields of chaotic iterations and PRNGs.

2.1 Notations

$[\![1;\mathsf{N}]\!]$ → $\{1, 2, \ldots, N\}$
S^n → the n^{th} term of a sequence $S = (S^1, S^2, \ldots)$
v_i → the i^{th} component of a vector $v = (v_1, v_2, \ldots, v_n)$
f^k → k^{th} composition of a function f
strategy → a sequence which elements belong in $[\![1;\mathsf{N}]\!]$
$\mathbb{S}$ → the set of all strategies
$\oplus$ → bitwise exclusive or
$+$ → the integer addition
$\ll$ and $\gg$→ the usual shift operators

2.2 Chaotic Iterations

Definition 1. *The set* $\mathbb{B}$ *denoting* $\{0, 1\}$*, let* $f : \mathbb{B}^{\mathsf{N}} \longrightarrow \mathbb{B}^{\mathsf{N}}$ *be an "iteration" function and* $S \in \mathbb{S}$ *be a chaotic strategy. Then, the so-called* chaotic iterations *are defined by [16]*

$$\begin{array}{l} x^0 \in \mathbb{B}^{\mathsf{N}}, \\ \forall n \in \mathbb{N}^*, \forall i \in [\![1;\mathsf{N}]\!], x_i^n = \begin{cases} x_i^{n-1} & if\, S^n \neq i \\ f(x^{n-1})_{S^n} & if\, S^n = i. \end{cases} \end{array} \tag{1}$$

In other words, at the n^{th} iteration, only the S^n-th cell is "iterated".

2.3 Input Sequences

In [17], we have designed a PRNG which has successfully passed the NIST tests suite. Unfortunately, this PRNG is too slow to pass the TestU01 battery of tests. Our ancient

PRNG which is called CI(Logistic, Logistic) PRNG is based on chaotic iterations and uses logistic maps as input sequences. However, chaotic systems like logistic maps work in the real numbers domain, and therefore a transformation from real numbers into integers is needed. This process leads to a degradation of the chaotic behavior of the generator and a lot of time wasted during computations. Moreover, a recent study shows that the use of logistic map for cryptographic applications is inadequate and must be discouraged [1]. Our purpose is then to design a new, faster, and more secure generator, which is able to pass the TestU01 battery of tests. This is achieved by using some faster PRNGs like ISAAC [9] and XORshift [13] as input sequences.

3 Design of CI(ISAAC,XORshift)

3.1 Chaotic Iterations as PRNG

The novel generator is designed by the following process. Let $\mathsf{N} \in \mathbb{N}^*, \mathsf{N} \geqslant 2$. Some chaotic iterations are fulfilled to generate a sequence $(x^n)_{n\in\mathbb{N}} \in \left(\mathbb{B}^{\mathsf{N}}\right)^{\mathbb{N}}$ of boolean vectors: the successive states of the iterated system. Some of these vectors are randomly extracted and their components constitute our pseudo-random bit flow. Chaotic iterations are realized as follows. Initial state $x^0 \in \mathbb{B}^{\mathsf{N}}$ is a boolean vector taken as a seed and chaotic strategy $(S^n)_{n\in\mathbb{N}} \in [\![1, \mathsf{N}]\!]^{\mathbb{N}}$ is constructed with XORshift. Lastly, iterate function f is the vectorial boolean negation

$$f_0 : (x_1, ..., x_{\mathsf{N}}) \in \mathbb{B}^{\mathsf{N}} \longmapsto (\overline{x_1}, ..., \overline{x_{\mathsf{N}}}) \in \mathbb{B}^{\mathsf{N}}.$$

To sum up, at each iteration only S^i-th component of state X^n is updated, as follows

$$x_i^n = \begin{cases} x_i^{n-1} \text{ if } i \neq S^i, \\ \\ \overline{x_i^{n-1}} \text{ if } i = S^i. \end{cases} \tag{2}$$

Finally, let $\mathcal{M}$ be a finite subset of $\mathbb{N}^*$. Some x^n are selected by a sequence m^n as the pseudo-random bit sequence of our generator. The sequence $(m^n)_{n\in\mathbb{N}} \in \mathcal{M}^{\mathbb{N}}$ is computed with ISAAC. So, the generator returns the following values: the components of x^{m^0}, followed by the components of $x^{m^0+m^1}$, followed by the components of $x^{m^0+m^1+m^2}$, *etc.* In other words, the generator returns the following bits:

$$x_1^{m_0} x_2^{m_0} x_3^{m_0} \dots x_{\mathsf{N}}^{m_0} x_1^{m_0+m_1} x_2^{m_0+m_1} \dots x_{\mathsf{N}}^{m_0+m_1} x_1^{m_0+m_1+m_2} x_2^{m_0+m_1+m_2} \dots$$

or the following integers:

$$x^{m_0} x^{m_0+m_1} x^{m_0+m_1+m_2} \dots$$

The basic design procedure of the novel generator is summed up in Table 1. The internal state is x, the output array is r. a and b are those computed by ISAAC and XORshift generators. Lastly, c and N are constants and $\mathcal{M}$ = {c, c+1} ($c \geqslant 3\mathsf{N}$ is recommended).

3.2 Example

In this example, $\mathsf{N} = 5$ and $\mathcal{M} = \{4,5\}$ are chosen for easy understanding. The initial state of the system x^0 can be seeded by the decimal part of the current time. For example,

```
Input: the internal state $x$ (an array of N bits)
Output: an array $r$ of N bits
$a \leftarrow ISAAC()$;
$m \leftarrow a \mod 2 + c$;
for $i = 0, \ldots, m$ do
    $b \leftarrow XORshift()$;
    $S \leftarrow b \mod \mathsf{N}$;
    $x_S \leftarrow \overline{x_S}$;
end
$r \leftarrow x$;
return $r$;
```

Algorithm 1. An arbitrary round of CI(ISAAC, XORshift)

Table 1. Application example

m :	4					5						4				
S	2	4	2	2		5	1	1	5	5		3	2	3	3	
In thisx^0					x^4						x^9					x^{13}
1					1		$\xrightarrow{1} 0$	$\xrightarrow{1} 1$			1					1
0	$\xrightarrow{2} 1$		$\xrightarrow{2} 0$	$\xrightarrow{2} 1$	1						1		$\xrightarrow{2} 0$			0
1					1						1	$\xrightarrow{3} 0$		$\xrightarrow{3} 1$	$\xrightarrow{3} 0$	0
0		$\xrightarrow{4} 1$			1						1					1
0					0	$\xrightarrow{5} 1$			$\xrightarrow{5} 0$	$\xrightarrow{5} 1$	1					1

Binary Output: $x_1^0x_2^0x_3^0x_4^0x_5^0x_1^4x_2^4x_3^4x_4^4x_5^4x_1^9x_2^9x_3^9x_4^9x_5^9x_1^{13}x_2^{13}...$ = 10100111101111110... Integer Output: $x^0, x^0, x^4, x^6, x^8...$ = 20, 30, 31, 19...

the current time in seconds since the Epoch is 1237632934.484084, so $t = 484084$. $x^0 = t \pmod{32}$ in binary digits, then $x^0 = (1,0,1,0,0)$. m and S can now be computed from ISAAC and XORshift:

- m = 4, 5, 4, 4, 4, 4, 5, 5, 5, 5, 4, 5, 4,...
- S = 2, 4, 2, 2, 5, 1, 1, 5, 5, 3, 2, 3, 3,...

Chaotic iterations are done with initial state x^0, vectorial logical negation f_0 and strategy S. The result is presented in Table 3. Let us recall that sequence m gives the states x^n to return: $x^4, x^{4+5}, x^{4+5+4}, \ldots$

So, in this example, the generated binary digits are: 10100111101111110011... Or the integers are: 20, 30, 31, 19...

3.3 Chaotic Iterations and Chaos

Generally the success of a PRNG depends, to a large extent, on the following criteria: uniformity, independence, storage efficiency, and reproducibility. A chaotic sequence may have these good pseudo-random criteria and also other chaotic properties, such as:

ergodicity, entropy, and expansivity. A chaotic sequence is extremely sensitive to the initial states. That is, even a minute difference in the initial state of the system can lead to enormous differences in the final state even over fairly small timescales. Therefore, chaotic sequence well fits the requirements of pseudo-random sequence. Contrary to ISAAC or XORshift, our generator possesses these chaotic properties.

However, despite a huge number of papers published in the field of chaos-based PRNGs, the impact of this research is rather marginal. This is due to the following reasons: almost all PRNG algorithms using chaos are based on dynamical systems defined on continuous sets (e.g., the set of real numbers). So these generators are usually slow, require considerably more storage spaces, and lose their chaotic properties during computations. These major problems restrict their use as generators [10]. Moreover, even if the algorithm obtained by the inclusion of chaotic maps is itself chaotic, the implementation of this algorithm on a machine can cause it lose its chaotic nature. This is due to the finite nature of the machine numbers set.

In this paper we don't simply integrate chaotic maps hoping that the implemented algorithm remains chaotic. The PRNG algorithms we conceive are constituted by discrete chaotic iterations that we mathematically proved in [3], that produce topological chaos as defined by Devaney. In the same paper, we raised the question of their implementation, proving in doing so that it is possible to design a chaotic algorithm and a chaotic computer program. In conclusion, the generator proposed in this paper does not inherit its chaotic properties from a continuous real chaotic map, but from discrete chaotic iterations defined in Section 2.2. As quoted above, it has been proven in [3] that chaotic iterations behave as chaos, as it is defined by Devaney: they are regular, transitive and sensitive to initial conditions. This famous definition of a chaotic behavior for a dynamical system implies unpredictability, mixture, sensitivity and uniform repartition. This allows the conception of a new generation of chaotic PRNGs. Because only integers are manipulated in discrete chaotic iterations, the chaotic behavior of the system is preserved during computations, and these computations are fast.

4 Security Analysis

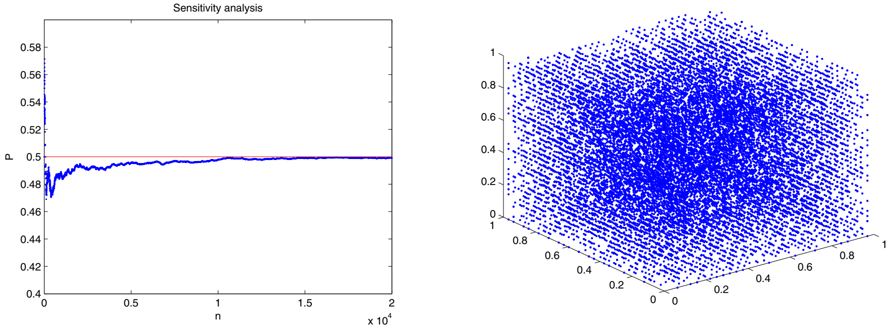
In this section a security analysis of the proposed generator is given.

4.1 Key Space

The PRNG proposed in this document is based on discrete chaotic iterations. It has an initial value $x^0 \in \mathbb{B}^{\mathsf{N}}$. Considering this set of initial values alone, the key space size is equal to 2^{N}. In addition, this PRNG combines digits of two other PRNGs: ISAAC and XORshift. Let k_1 and k_2 be the key spaces of ISAAC and XORshift. So the total key space size is close to $2^{\mathsf{N}} \cdot k_1 \cdot k_2$. Finally, the impact of $\mathcal{M}$ must be taken into account. This leads to conclude that the key space size is large enough to withstand attacks.

4.2 Key Sensitivity

This PRNG is highly sensitive to the initial conditions. To illustrate this property proved in [3], several initial values are put into the chaotic system. Let H be the number of



a. sensitivity

b. Second order distribution

Fig. 1. Security analysis

differences between the sequences obtained in this way. Suppose n is the length of these sequences. Then the variance ratio P, defined by $P = H/n$, is computed. The results are shown in Figure 1a (x axis is sequence lengths, y axis is variance ratio P). Variance ratios approach 0.50, which indicates that the system is extremely sensitive to the initial conditions.

4.3 Uniform Distribution

Figure 1b gives a 3D graphic representation of the distribution of a random sequence obtained by our generator. The point cloud presents a uniform distribution that tends to fill the complete 3D space, as expected for a random signal. To obtain this cloud, we have first changed the binary sequence to a N-bit integer sequence x_1, x_2, x_3, x_4... Then we have plot $\left(\frac{x_1}{2^N}, \frac{x_2}{2^N}, \frac{x_3}{2^N}\right), \left(\frac{x_2}{2^N}, \frac{x_3}{2^N}, \frac{x_4}{2^N}\right)$...

5 TestU01 Statistical Test Results

In a previous section, we have shown that the proposed PRNG has strong chaotic properties, as Devaney's chaos. In particular, this generator is better than the well-known XORshift and ISAAC, in the topological point of view. In addition to being chaotic, we will show in this section that CI(ISAAC,XORshift) is better than XORshift, and at least as good as ISAAC [18] in the statistical point of view. Indeed, similarly to ISAAC and contrary to XORshift, CI(ISAAC,XORshift) can pass the stringent Big Crush battery of tests included in TestU01. In addition, our generator achieves to pass all the batteries included in TestU01. To our best knowledge, this result has not been proven for ISAAC, and only one other generator is capable of doing this [6].

5.1 TestU01

Indeed, the quality of a PRNG should be based on theoretical fundamentals but should also be tested empirically. Various statistical tests are available in the literature that

Table 2. TestU01 Statistical Test

Battery	Parameters	Statistics
Rabbit	32×10^9 bits	40
Alphabit	32×10^9 bits	17
Pseudo DieHARD	Standard	126
FIPS_140_2	Standard	16
Small Crush	Standard	15
Crush	Standard	144
Big Crush	Standard	160

test a given sequence for some level of computational indistinguishability. Major test suites for RNGs are TestU01 [11], the NIST suite [15], and the DieHARD suites [12]. The DieHARD suites, which implement many classical RNG tests, have some drawbacks and limitations. The National Institute of Standards and Technology (NIST), in the United States, has implemented a test suite (16 tests) for RNGs. It is geared mainly for the testing and certification of RNGs used in cryptographic applications. TestU01 is extremely diverse in implementing classical tests, cryptographic tests, new tests proposed in the literature, and original tests. In fact, it encompasses most of the other test suites. The proposed PRNG has been tested using TestU01 for its statistical pseudo randomness.

5.2 Batteries of Tests

Table 2 lists seven batteries of tests in the TestU01 package. ”Standard” parameter in this Table refers to the built-in parameters of the battery. TestU01 suite implements 518 tests and reports $p-$values. If a $p-$value is within $[0.001, 0.999]$, the associated test is a success. A $p-$value lying outside this boundary means that its test has failed.

5.3 Analysis

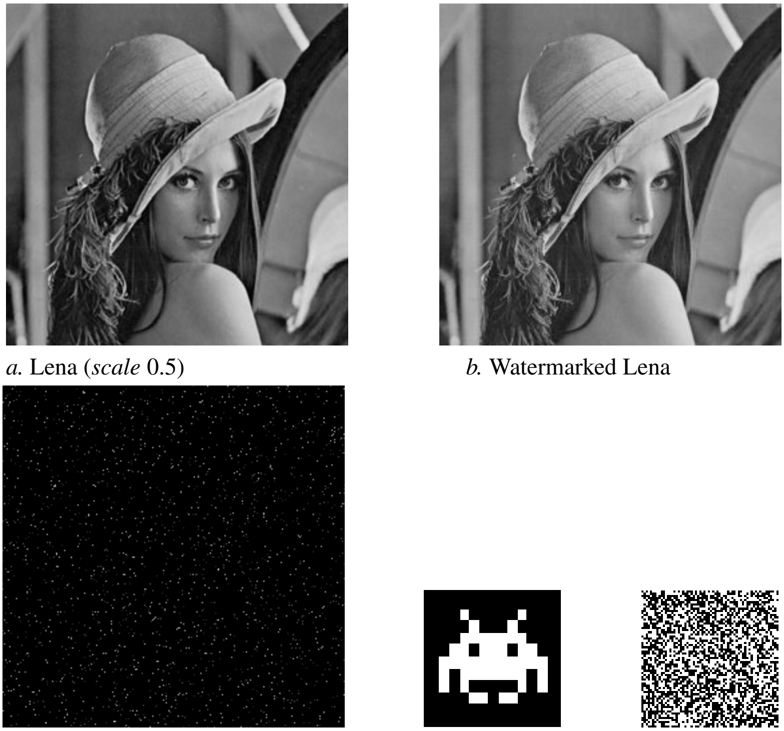
In a sound theoretical basis, a PRNG based on discrete chaotic iterations (ICs) is a composite generator which combines the features of two PRNGs. The first generator constitutes the initial condition of the chaotic dynamical system. The second generator randomly chooses which outputs of the chaotic system must be returned. The intention of this combination is to cumulate the effects of chaotic and random behaviors, to improve the statistical and security properties relative to each generator taken alone.

This PRNG based on discrete chaotic iterations may utilize any reasonable RNG as inputs. For demonstration purposes, XORshift and ISAAC are adopted here. The PRNG with these inputs can pass all of the performed tests.

6 Application Example in Digital Watermarking

In this section, an application example is given in the field of digital watermarking: a watermark is encrypted and embedded into a cover image using chaotic iterations and our PRNG. The carrier image is the famous Lena, which is a 256 grayscale image, and the watermark is the 64×64 pixels binary image depicted in Fig.2d. Let us encrypt the watermark by using chaotic iterations. The initial state x^0 of the system is constituted by the watermark, considered as a boolean vector. The iteration function is the vectorial logical negation f_0. The PRNG presented previously is used to obtain a sequence of integers lower than 4096, which will constitute the chaotic strategy $(S^k)_{k \in \mathbb{N}}$. Thus, the encrypted watermark is the last boolean vector generated by the chaotic iterations. An example of such an encryption, with 5000 iterations, is given in Fig.2e.

Let L be the 256^3 booleans vector constituted by the three last bits of each pixel of Lena. We define U^k by $U^0 = S^0$ and $U^{n+1} = S^{n+1} + 2 \times U^n + n\ [mod\ 256^3]$. The watermarked Lena I_w is obtained from the original Lena I_o, the three last bits of which are replaced by the result of 64^2 chaotic iterations with initial state L, and strategy U^k (see Fig.2b). Spatial domain embedding has been chosen here for easy understanding,



a. Lena (*scale* 0.5)

b. Watermarked Lena

c. Differences

d. Watermark

e. Encrypted watermark

Fig. 2. Original and watermarked Lena

but this watermarking scheme can be adapted to frequency domain (for an example of its use in DWT domain, see [2]). The extraction of the watermark can be obtained in the same way [2]. Remark that the map $\theta \mapsto 2\theta$ of the torus, which is the well-known dyadic transformation (an example of topological chaos [7]), has been chosen to make $(U^k)_{k\leqslant 64^2}$ highly sensitive to the chaotic encryption strategy.

The robustness of this data hiding scheme through geometric and frequency attacks has been studied in [2]. The chaos-security and stego-security are proven in [8]. The difference with the scheme presented in these papers is the way to generate strategies, *i.e.*, the choice of the initial conditions for chaotic iterations, in the encryption and embedding stages. This improvement does not alter robustness and subspace-security. We have shown in this study that this replacement enhances the speed of the scheme. Moreover, it resolves a potential security lack related to the use of a logistic map [1] when generating the strategies: this lack might be exploited by an attacker in Watermark-Only-Attack and Known-Message-Attack setups [4]. Instead of logistic map, our PRNG has good statistical properties and can withstand such attacks. This claim will be deepened in future work.

7 Conclusions and Future Work

In this paper, the PRNG proposed in [17] is improved. This is achieved by using the famous ISAAC and XORshift generators and by combining these components with chaotic iterations. Thus we obtain a faster generator which satisfies chaotic properties. In addition to passing the NIST tests suite, this new generator successfully passes all the stringent TestU01 battery of tests. The randomness and disorder generated by this algorithm has been evaluated. It offers a sufficient level of security for a whole range of applications in computer science. An application example in the field of data hiding is finally given. In future work, the comparison of different chaotic strategies will be explored and other iteration functions will be studied. Finally, other applications in computer science security field will be proposed, especially in cryptographic domains.

References

1. Arroyo, D., Alvarez, G., Fernandez, V.: On the inadequacy of the logistic map for cryptographic applications. X Reniun Espaola Sobre Criptologa y Seguirdad de la Informacin (XRECSI) 1, 77–82 (2008)
2. Bahi, J., Guyeux, C.: A new chaos-based watermarking algorithm. In: SECRYPT 2010, International conference on security and cryptography, Athens, Greece, pp.35-40 (to appear, 2010)
3. Bahi, J.M., Guyeux, C.: Chaotic iterations and topological chaos. arXiv 0810.3154 (2008)
4. Cayre, F., Bas, P.: Kerckhoffs-based embedding security classes for woa data hiding. IEEE Transactions on Information Forensics and Security 3(1), 1–15 (2008)
5. Cecen, S., Demirer, R.M., Bayrak, C.: A new hybrid nonlinear congruential number generator based on higher functional power of logistic maps. Chaos, Solitons and Fractals 42, 847–853 (2009)
6. Corsaro, S., De Angelis, P.L., Marino, Z., Perla, F., Zanetti, P.: On parallel asset-liability management in life insurance: a forward risk-neutral approach. Parallel Computing (2009)(in Press)

7. Devaney, R.L.: An Introduction to Chaotic Dynamical Systems, 2nd edn. Addison-Wesley, Reading (1989)
8. Guyeux, C., Friot, N., Bahi, J.M.: A more secure information hiding scheme than spread-spectrum obtained by chaos-security. arXiv 0032565 (2010)
9. Jenkins, R.J.: Isaac. Fast Software Encryption, 41–49 (1996)
10. Kocarev, L.: Chaos-based cryptography: a brief overview. IEEE Circ Syst Mag 7, 6–21 (2001)
11. L'ecuyer, P., Simard, R.: Testu01: A software library in ansi c for empirical testing of random number generators. Laboratoire de simulation et doptimisation. Universi de Montral IRO (2009)
12. Marsaglia, G.: Diehard: a battery of tests of randomness (1996), http://stat.fsu.edu/~geo/diehard.html
13. Marsaglia, G.: Xorshift rngs. Journal of Statistical Software 8(14), 1–6 (2003)
14. Po-Han, L., Yi, C., Soo-Chang, P., Yih-Yuh, C.: Evidence of the correlation between positive lyapunov exponents and good chaotic random number sequences. Computer Physics Communications 160, 187–203 (2004)
15. NIST Special Publication 800-22 rev. 1. A statistical test suite for random and pseudorandom number generators for cryptographic applications (August. 2008)
16. Robert, F.: Discrete Iterations. A Metric Study. Springer Series in Computational Mathematics, vol. 6. Springer, Heidelberg (1986)
17. Wang, Q., Guyeux, C., Bahi, J.M.: A novel pseudo-random generator based on discrete chaotic iterations for cryptographic applications. In: First International Conference on Evolving Internet (2009)
18. Wichmanna, B.A., Hillb, I.D.: Generating good pseudo-random numbers. Computational Statistics & Data Analysis 51, 1614–1622 (2006)
19. Zheng, F., Tian, X., Song, J., Li, X.: Pseudo-random sequence generator based on the generalized henon map. The Journal of China Universities of Posts and Telecommunications 15(3), 64–68 (2008)

A Secure Protocol for Point-Segment Position Problem

Yi Sun[1], Hongxiang Sun[2], Hua Zhang[1], and Qiaoyan Wen[1]

[1] State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, 100876, China
sunyi__sunyi@163.com, {zhanghua_288,wqy}@bupt.edu.cn
[2] School of Science, Beijing University of Posts and Telecommunications,
Beijing, 100876, China
shx@bupt.edu.cn

Abstract. Privacy Preserving Computation Geometry is an important direction in the application of Secure Multi-party Computation and contains many research subjects, such as intersection problem, point-inclusion problem, convex hull, rang searching and so on. Particularly, point-inclusion problem is of great practical significance in our daily life. In this paper, we will devote our attention to the point-segment position problem in point-inclusion and aim to determine the relationship of a point and a segment. In our solution, we present a concise secure protocol based on two basic protocols, secure scalar product protocol and secure comparison protocol. Compared with precious solutions, which may disclose at least one inside point, our protocol performs better in terms of preserving privacy. It will not reveal any inside point, which is crucially significant in some special occasion.

Keywords: secure multi-party computation; point-inclusion problem; straddled segments; privacy preserving computation geometry; point-segment position problem.

1 Introduction

Secure Multi-party Computation (SMC) is dedicated to deal with the problem of secure computation among distrustful participants. It was first introduced by Yao in 1982 [1], and then was extended by Goldreich, Micali Wigderson [2] and many other researchers [3, 4, 5]. Generally speaking, SMC is a method to implement cooperative computation with participants' private data, ensuring the correctness of the computation as well as not disclosing additional information except the necessary results. Until now, SMC has become a research focus in the international cryptographic community. Secure two-party computation (STC) [3, 6, 7] is a special case in SMC since there are only two participants in the process. The well-known millionaires' problem put forward by Yao [1] is the representative problem in STC. In our discussing, we will consider the two-party case.

Privacy Preserving Computation Geometry (PPCG), a classic realm in SMC, was first brought forward by Mikhail J Atllah and Wenliang Du [7]. Due to its great practical significance, great efforts have been poured into this field since then.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 212–219, 2010.

Nowadays, it has been widely applied to various scenarios like commerce, military, government and so on [3]. In this field, Point-inclusion problem [8, 9, 10] is a large branch of PPCG. The core issue of point-inclusion is to determine whether a certain point is in the related area. Usually, the solution to this problem is first to determine the relationship of the point and one side of the area, and then decide if the point is in the area. In [11], the users execute the protocol with the help of the original point *O*. In case *O* is not inside of the area, it is unavoidable to disclose an inside point of the area to the other party in the initialization phase. However, in some special occasion, we are not allowed to leak any inside information, even if only one point. In order to protect all inside points, we present a concise secure protocol based on secure scalar product protocol (SPP) and secure comparison protocol (SCP) to judge the relationship of a point and a segment--that is what we called point-segment position problem in this paper.

Compared with precious solutions, which may disclose at least one inside point, our protocol performs better in terms of security. Our main improvement is to eliminate the dependence of the original point *O*. In this way, our protocol will not reveal any inside point, which is crucially significant in some special occasion.

The rest of the paper is organized as follows. Section II introduces some preliminaries used in our protocol. Then, we present our protocol and analyze the correctness, security and complexity in section III. Finally, in section IV, we conclude by summarizing our results and pointing out some significant issues in the future work.

2 Preliminaries

In this section, we give an over view of the main framework of two basic protocols in SMC (SPP and SCP), one computational formula of triangle area, and a definition of straddled segments.

Denoting the two participants as Alice and Bob separately.

SPP is a primary and fundamental protocol in SMC, especially in the scope of PPCG. It was first proposed in [7], and developed rapidly since then [12, 13, 14]. The main process is described as follows. We denote the communication complexity and time complexity of the invoked SPP as C_{SPP} and T_{SPP} separately.

Secure Scalar Product Protocol (SPP)

Input: Alice has a vector $X = (x_1, \cdots, x_n)$; Bob has a vector $Y = (y_1, \cdots, y_n)$.

Output: Alice gets V, which is randomly chosen by itself, and keeps V as its private data; Bob gets U, where $U = V + \sum_{i=1}^{n} x_i y_i$. Bob keeps U as its private data.

SCP is a typical two-party protocol in SMC, which has solved the well-know millionaires' problem brought out by Yao in 1982 [1]. Until now, there have been many variant forms of the primary SCP [15, 16, 17]. We denote the communication complexity and time complexity of the invoked SCP as C_{SPP} and T_{SPP} separately.

Secure Comparison Protocol (SCP)

Input: Alice has a private data V; Bob has a private data U.

Output: Alice and Bob get the value of U-V, which implies: U>V, if U-V>0; or U<V, if U-V<0; or U=V, if U-V=0.

Area of a Triangle: There is a triangle $P_1P_2Q_1$. The co-ordinates of its three vertexes P_1, P_2 and Q_1 are (x_{P_1}, y_{P_1}), (x_{P_2}, y_{P_2}) and (x_{Q_1}, y_{Q_1}) respectively. Then the area of the triangle $P_1P_2Q_1$ can be easily calculated by the following formula:

$$S_{\Delta P_1P_2Q_1} = [(x_{P_1}y_{P_2} - x_{P_2}y_{P_1}) + y_{Q_1}(x_{P_2} - x_{P_1}) + x_{Q_1}(y_{P_1} - y_{P_2})] / 2$$

Straddled Segments: Segment P_1P_2 is straddled with segment Q_1Q_2 if and only if $(P_1 - Q_1)\times(Q_2 - Q_1) * (Q_2 - Q_1)\times(P_2 - Q_1) \ge 0$, Where $P_1 \times P_2 = x_{P_1} * y_{P_2} - x_{P_2} * y_{P_1}$.

3 Point-Segment Position Problem

3.1 Point-Segment Position Problem

The concrete problem we want to solve can be described as follows. There are two straddled segments P_1P_2 and Q_1Q_2, which belongs to Alice and Bob separately. As we know, two straddled segments have the following three cases according to the relation of point and segment in figure 1.

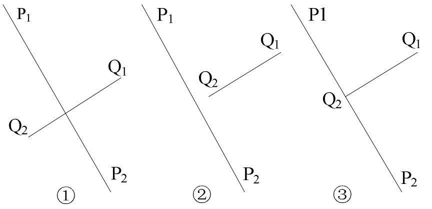


Fig. 1. The Relation of Point and Segment

Our task is to determine the position of Q_2: on the left side of P_1P_2 or the right side of P_1P_2, or just on this segment P_1P_2 (Here, without loss of generality, we only concern whether Q_1 and Q_2 are on the same side of P_1P_2 or on the two sides of P_1P_2). To put forward the point-segment position problem formally, we have

Point-Segment Position Problem

Input: Alice has a private segment P_1P_2; Bob has a private segment Q_1Q_2.

Output: Alice and Bob get the position relation between Q_2 and segment P_1P_2.

This problem is really practical in our daily life especially in commerce. Next, we will provide the point-segment position problem a typical application scenario in commercial area. Suppose two rivalrous companies Alice and Bob plan to explore a new market in a business street. In order to obtain the maximum interests, both of them would like to have a cooperative evaluation about their plans on the premise of keeping their privacy. Figure 2 helps us to depict the above problem.

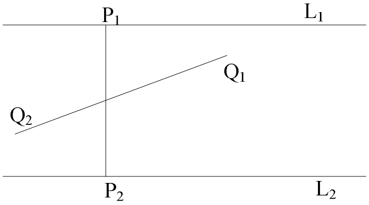


Fig. 2. A Practical Problem

There is a business street denoted as L_1L_2. In Alice's plan, P_1P_2 is the left influenced boundary of its new market.(Without loss of generality, we can assume that P_1P_2 is vertical with L_1L_2) Q_2 is the point where Bob want to open a new shop, Q_1 is another point on the right side of Q_2, which is randomly chosen by Bob as an auxiliary point. Both Alice and Bob want to know Q_2 is on which side of P_1P_2.

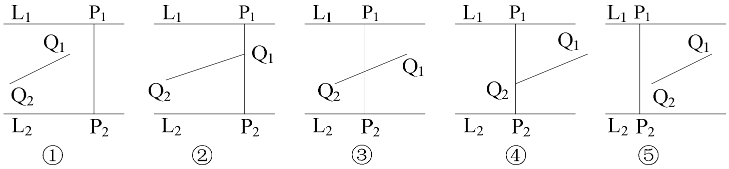


Fig. 3. The five cases

Combining with the application example, we can see that it is reasonable to divide the whole instance into three cases as described in figure 1. In the application circumstance, according to P_1P_2, we get five cases as described in figure 3. But essentially, case ③ is the same situation as the first case ① in figure 1, that is, Q_2 is on the left side of segment P_1P_2; Similarly, case ④ in figure 3 is just as the same as the third case ③ in figure 1; case ⑤ in figure 3 is as the same as the second case ② in figure 1. For case ① and ② in figure 3, because Q_2 is the point where Bob wants to open a new shop and Q_1 is another point on the right side of Q_2, which is randomly chosen by Bob as an auxiliary point. We can limit Bob to choose the point Q_1 at the boundary of the business street so that case ① and ② in figure 3 are included in the following three cases. In this way, the problem in our daily life is reduced to the point-segment position problem.

3.2 Secure Point-Segment Position Protocol

Now, we analyze the problem and present point-segment position protocol (PSPP) to solve the above problem.

In a nutshell, we will distinguish the three possible cases depending on the area relationship of related regions. It is easy to have the following three equivalent propositions:

a. "Q_2 is on the left side of segment P_1P_2" is equal to "the area of triangle $P_1Q_1Q_2$ and triangle $P_2Q_1Q_2$ is larger than the area of triangle $P_1P_2Q_1$";

b. "Q_2 is on the right side of segment P_1P_2" is equal to "the area of triangle $P_1Q_1Q_2$ and triangle $P_2Q_1Q_2$ is smaller than the area of triangle $P_1P_2Q_1$";

c. "Q_2 is on the segment P_1P_2" is equal to "the area of triangle $P_1Q_1Q_2$ and triangle $P_2Q_1Q_2$ is equal to the area of triangle $P_1P_2Q_1$".

In order to simulate the application circumstance in our abstract discussing and meet the requirements of the problem defined before, we first need to do some initializations.

a. Alice and Bob agree on a common execute region.

From P_1P_2, we can get two parallel lines L_1 and L_2, which is vertical with P_1P_2 at P_1 and P_2 separately. The region between L_1L_2 is what we want in the execution.

b. Bob randomly choose a point Q_1 at the boundary of the business street on the right side of Q_2 as its auxiliary point in the common region.

In this way, we can guarantee that P_1P_2 and Q_1Q_2 are two straddled segments. At the same time, it coincides with the application scenario. According to the formula of triangle area mentioned in section 2, we have,

$$S_{\Delta P_1P_2Q_1} = [(x_{P_1}y_{P_2} - x_{P_2}y_{P_1}) + y_{Q_1}(x_{P_2} - x_{P_1}) + x_{Q_1}(y_{P_1} - y_{P_2})]/2$$

$$S_{\Delta Q_1Q_2P_1} = [(x_{Q_2}y_{Q_1} - x_{Q_1}y_{Q_2}) + x_{P_1}y_{Q_2} - x_{P_1}y_{Q_1} + x_{Q_1}y_{P_1} - x_{Q_2}y_{P_1}]/2$$

$$S_{\Delta Q_1Q_2P_2} = [(x_{Q_2}y_{Q_1} - x_{Q_1}y_{Q_2}) + x_{P_2}y_{Q_2} - x_{P_2}y_{Q_1} + x_{Q_1}y_{P_2} - x_{Q_2}y_{P_2}]/2$$

Now, for our problem we only need to discuss the value:

$$S_{\Delta Q_1Q_2P_2} + S_{\Delta Q_1Q_2P_1} - S_{\Delta P_1P_2Q_1} = (x_{Q_1}y_{P_2} - x_{P_2}y_{Q_1}) + [y_{Q_2}(x_{P_1} + x_{P_2})$$
$$-x_{Q_2}(y_{P_1} + y_{P_2})]/2 + (x_{Q_2}y_{Q_1} - x_{Q_1}y_{Q_2}) - (x_{P_1}y_{P_2} - x_{P_2}y_{P_1})/2$$

Denote $\sigma_1 = x_{Q_1}y_{P_2} - x_{P_2}y_{Q_1}$, $\sigma_2 = [y_{Q_2}(x_{P_1} + x_{P_2}) - x_{Q_2}(y_{P_1} + y_{P_2})]/2$,

$\rho_U = x_{Q_2}y_{Q_1} - x_{Q_1}y_{Q_2}$, $\rho_V = (x_{P_1}y_{P_2} - x_{P_2}y_{P_1})/2$

We have,

$$S_{\Delta Q_1Q_2P_2} + S_{\Delta Q_1Q_2P_1} - S_{\Delta P_1P_2Q_1} = \sigma_1 + \sigma_2 + \rho_U - \rho_V$$

Inspired by this expression, we can present our point-segment position protocol (PSPP) based on SPP and SCP:

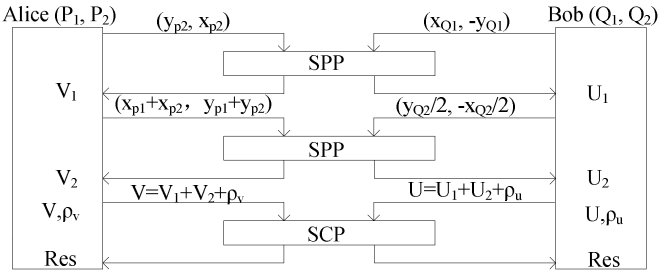


Fig. 4. Point-Segment Position Protocol

Point-Segment Position Protocol (PSPP)

Input: Alice has a private segment P_1P_2; Bob has a private segment Q_1Q_2.

Output: Alice and Bob get the relation between point Q_2 and segment P_1P_2: Q_2 is on the left side of segment P_1P_2 or on the right side of P_1P_2 or is just on the segment P_1P_2.

Step 1: Invoking SPP

Input: Alice inputs $(\mathrm{y}_{\mathrm{P}_2}, \mathrm{x}_{\mathrm{P}_2})$; Bob inputs $(x_{Q_1}, -y_{Q_1})$.

Output: Alice gets V_1, which is randomly chosen by itself and keeps it as its secret;

Bob gets U_1, where $U_1 = V_1 + \sigma_1$, $\sigma_1 = x_{Q_1} y_{P_2} - x_{P_2} y_{Q_1}$. Bob keeps U_1 as its secret.

Step 2: Invoking SPP

Input: Alice inputs $(x_{P_1} + x_{P_2}, y_{P_1} + y_{P_2})$; Bob inputs $(y_{Q_2} / 2, -x_{Q_2} / 2)$.

Output: Alice gets V_2, which is randomly chosen by itself and keeps it as its secret;

Bob gets U_2, where $\sigma_2 = [y_{Q_2}(x_{P_1} + x_{P_2}) - x_{Q_2}(y_{P_1} + y_{P_2})] / 2$, $U_2 = V_2 + \sigma_2$. Bob keeps U_2 as its secret.

Step 3: Computing locally

Alice computes $\rho_V = (x_{P_1} y_{P_2} - x_{P_2} y_{P_1}) / 2, V = V_1 + V_2 + \rho_V$ locally;

Bob computes $\rho_U = x_{Q_2} y_{Q_1} - x_{Q_1} y_{Q_2}, U = U_1 + U_2 + \rho_U$ locally.

Step 4: Invoking SCP

Input: Alice inputs V; Bob inputs U.

Output: Alice and Bob get the value of U-V.

To prevent revealing the accurate value of U-V, we can play a little trick. During executing the secure comparison protocol in step 4, we generate a variable which is denoted as Res: If U-V>0, we set Res =1; If U-V<0, we set Res=-1; If U-V=0, we set Res =0. By this trick, we have the new form of output. That is,

Output: Alice and Bob get the value of Res.

3.3 Analysis of the Protocol

Secure multi-party protocols are often analyzed in correctness, security, and complexity. In terms of correctness, we always compare the targeted protocol to the ideal one where there is a trusted third party (TTP) to help participants perform the cooperative computation. This property ensures that each party can achieve the desirable results. In security, we must guarantee that the privacy of each party has not learned by the other one, which is the heart of SMC. Finally, in complexity, we usually show two indexes, computational and time complexity.

Correctness analysis

The correctness of point-segment position protocol (PSPP) is obvious, which is implied in the foregoing discussing of our protocol. According to the previous deduction of the relationship between our problem and the area of related triangles, we can easily find that PSPP is correct indeed.

Security analysis

The security of step 1 and step 2 are guaranteed by secure scalar product protocol (SPP). Correspondingly, the security of step 4 is based on the security of secure comparison protocol (SCP). And the security of step 3 is doubtless since there is no interaction between Alice and Bob. Overall, we can see that neither Alice nor Bob can gain additional information except the necessary result. Thus, PSPP is secure.

Complexity analysis

In our protocol, we might as well regard the invoked basic protocols as sub-protocols. Here the complexity of point-segment position protocol (PSPP) denoted as C_{PSPP} and T_{PSPP} are described by the complexity of SPP and SCP. Hence the computational complexity C_{PSPP} is $o(2C_{SPP} + C_{SCP})$, and the time complexity of PSPP T_{PSPP} is $o(2T_{SPP} + T_{SCP})$.

4 Conclusions and Future Work

In this paper, we have addressed the problem of privacy preserving point-segment position, which is a crucial aspect in PPCG, and being widely applied to various fields in our daily life, such as commerce, military, government and so on. There has been a concise secure protocol presented in this paper to solve this problem based on SPP and SCP in order to make our protocol more attractive to users in the real life. The corresponding analysis of correctness, security and complexity are also presented.

Though our protocol is capable of solving the problem, there still leaves some spaces to perfect our solution. Extending it to multi-party case is our next task in the future.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant Nos. 60873191, 60903152, 60821001) and Beijing Natural Science Foundation (Grant No. 4072020).

References

1. Yao, A.C.: Protocols for secure computations. In: Proceedings of the 23th Annual IEEE Symposium on Foundations of Computer Science, pp. 160–164. IEEE Computer Society Press, Los Alamitos (1982)
2. Goldreich, O., Micali, S., Wigderson, A.: How to play mental game. In: Proceedings of the 19th ACM Symposium on Theory of Computing (STOC), pp. 218–229. ACM Press, New York (1987)
3. Du, W., Atallah, M.J.: Secure multi-party computation problems and their applications: A review and open problems. In: Proceedings of New Security Paradigms Workshop, Cloudcroft, New Mexico, USA, pp. 11–20 (2001)
4. Goldreich, O.: Secure multi-party computation. (manuscript version 1.3) (2002), `http://theory.lcs.mit.edu/~oded`
5. Luo, Y., Huang, L., Zhong, H., et al.: A Secure protocol for determining whether a point is inside a convex polygon. Chinese Journal of Electronics 15(4), 578–582 (2006)
6. Lindell, Y.: Parallel coin-tossing and constant-round secure two-party computation. Journal of Cryptology 16(3), 143–184 (2003)
7. Atllah, M.J., Du, W.: Secure multi-party computation on geometry. In: Dehne, F., Sack, J.-R., Tamassia, R. (eds.) WADS 2001. LNCS, vol. 2125, pp. 165–179. Springer, Heidelberg (2001)
8. Ramon, J., Katzenbesser, S.: A secure multimensional point inclusion protocol. In: MM&Sec 2007, Texas, USA, pp. 109–120. ACM, New York (2007)
9. Sang, Y., Shen, H.: A scheme for testing privacy state in perasive sensor network. In: NINA 2005 (2005)
10. Hui, K.C.: A robust point inclusion algorithm regions bounded by parametric curve segments. Computer-aided design 29, 771–778 (1997)
11. Ye, Y., Huang, L., Yang, W., Zhou, Z.: Efficient secure protocol to determine whether a point is inside a convex hull. In: IEEE International Symposium on Information Engineering and Electronic Commerce, IEEC 2009, pp. 100–105. IEEE Press, Los Alamitos (2009)
12. Amirbekyan, A., Estivill-Castro, V.: A new efficient private preserving scalar product protocol. In: 6th Australasian Data Mining Conference (AusDM 2007), pp. 209–214 (2007)
13. Du, W., Zhan, Z.: Building decision tree lassifier on private data. In: IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining, pp. 1–8 (2002)
14. Wang, I.-C., Shen, C., et al.: Towards empirical aspects of secure scalar product. In: International Conference on Information Security and Assurance, pp. 573–578 (2008)
15. Cachin, C.: Efficient private bidding and auctions with an obvious third party. In: Proceedings of the 6th ACM Conference on Computer and Communications Security, pp. 120–127. ACM Press, New York (1999)
16. Ioannidis, I., Grama, A.: An efficient protocol for Yao's millionaires' problem. In: Proceedings of the 36th Hawaii International Conference on System Sciences, HICSS 2003 (2003)
17. Lin, H.Y., Tzeng, W.G.: An efficient solution to the millionaires' problem based on hormonoorphic encryption. In: Ioannidis, J., Keromytis, A.D., Yung, M. (eds.) ACNS 2005. LNCS, vol. 3531, pp. 456–466. Springer, Heidelberg (2005)

Learning Automata Representation of Network Protocol by Grammar Induction

Ming-Ming Xiao[1,2] and Shun-Zheng Yu[1]

[1] Department of Electronics and Communication Engineering, Sun Yat-Sen University,
510006 Guangzhou, China
[2] Information College, Zhongkai University of Agriculture and Engineering,
510225 Guangzhou, China
xmingm@gmail.com, syu@mail.sysu.edu.cn

Abstract. In this work, the grammatical inference was applied to model network protocol specification as FSM from the network stream data. The original RPNI algorithm merges pairs of states of the prefix tree acceptor of the positive samples in a fixed order assuring consistency of the resulting automaton, which would get a over-generalization automaton. The proposals presented consist in the modification of RPNI algorithm by means of introducing heuristics about network feature that label merging states from the prefix tree acceptor to prevent state from merging excessively. Preliminary experiments done seem to show that the improvement over the original RPNI algorithm is more helpful for deriving the more general network protocol automaton.

Keywords: Automaton inference; Machine learning; Protocol analysis.

1 Introduction

It is well-known to be valuable about protocol reverse engineering for many network security applications, including intrusion prevention and detection that performs deep packet inspection and traffic normalization, penetration testing that generates network inputs to an application to uncover potential vulnerabilities, and so on. Unfortunately, current bread-and-butter approach to determine protocol specification depends on analyzing manually. Current technique of protocol reverse engineering rests on the stage of message extraction automatically. Whereas regular expression(RE) was extensively applied to newly system, such as l7-filter, Bro, Snort, and so on, RE also are generated manually based on analyzing the implementing context of application protocol in advance. So, we can conclude that there is not an effect method to model network protocol specification.

In artificial intelligence, linguistics and in other fields it is often important to find some regulations in given input data consisting of a subset of an unknown language (so called positive samples) and a subset of its complement (so called negative samples). If the regularity has to be represented as a finite state automaton (FSA) the problem is called regular inference [1]. From as early as in the fifties [2], Grammatical Inference was already conceived as a technique used to learn syntax from example sentences. Most of the research efforts are concentrated on the learning

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 220–227, 2010.

of finite automata (i.e. regular grammars) [3]. It has been used so far to plenty of fields extensively, and also been applied to information extraction.

In this work, we introduced the grammatical inference to model protocol specification as FSM from the extracted network trace. We propose a method using RPNI algorithm as a base and a labeled search algorithm to improve its performance through selecting the right set of pairs of states that should not be merged during the run of the original RPNI algorithm.

2 Grammar Inference

Grammar induction is a particular case of inductive learning. The processing architecture is shown in Fig. 1. The general law is represented by a formal grammar or an equivalent machine.

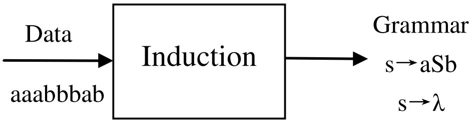


Fig. 1. Grammar induction overview

The set of examples, known as positive sample, is usually made of strings or sequences over a specific alphabet. A negative sample, i.e. a set of strings not belonging to the target language, can sometimes help the induction process. The positive data can be represented by a prefix tree acceptor (PTA). The PTA of example:{ aa, abba, baa} is shown in Fig. 2.

We observe some positive and negative data. The positive sample S^+ comes from a regular language L_0, The positive sample is assumed to be structurally complete with respect to the canonical automaton $A(L_0)$ of the target language L_0. We build the Prefix Tree Acceptor of S^+ By construction $L(PTA) = S^+$. The negative sample S^- helps to control over-generalization of the inferred automaton.

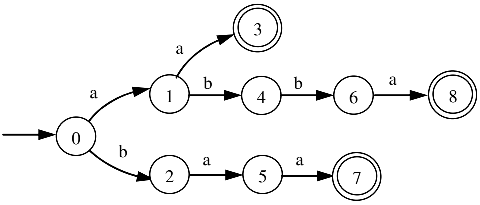


Fig. 2. The example of PTA

3 Approach and System Realization Design

In this section, we describe the detailed technique of our solution. We introduce the GI algorithms to learn the DFA for modeling the protocol application. We first present the original RPNI algorithm briefly, and introduce the overview of the system architecture. Then propose the heuristics of network to modify the RPNI algorithm, which is based on the method that the merging of state is controlled by labeling the state with difference.

3.1 RPNI Algorithm

A well-known RPNI (Regular Positive and Negative Inference) algorithm for regular language inference [4] has been proven to identify the target FSA and run in polynomial time under some additional conditions on the given samples. It can be briefly described as follows (algorithm taken from [5]):

At first a prefix tree acceptor PTA(S^+) for the samples from S^+ is constructed. Then the states of PTA(S^+) are numbered according to the standard order of the set Pr(S^+) (shorter strings precede the longer ones, strings of the same length are ordered lexicographically) as 0, . . .,N-1. The space of partitions of the set of states of PTA(S^+) (and thus the space of FSA's) is searched under the control of S^-. The initial partition (in the following code labeled π_0) corresponding to the PTA(S^+) itself is {{0}, {1}, . . . , {N-1}}. The blocks of the partition are iteratively merged while keeping the consistency of the corresponding quotient automaton with negative samples S^-. This algorithm is guaranteed to find a finite state automaton consistent with the given set of samples. On the other hand, RPNI is not guaranteed to find the smallest FSA consistent with the given set of samples.

3.2 System Architecture Design

The system architecture in our design is shown in Fig. 3. As depicted in the figure, our system design mainly consists of four stages in summary. The first stage is the message format extraction, which is implemented by execution monitor [6] and messages format extraction model [7]. It takes as input the program's binary and the application data, and dynamically monitors how the program processes the application data. The output of the execution monitor is an execution trace that contains a record of all the instructions performed by the program. And then the execution trace is analyzed to indicate the field boundaries and the keywords extraction during message format extraction. The result of the stage is message format sequence.

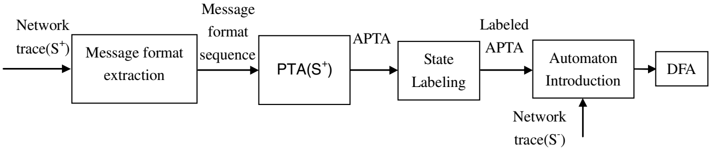


Fig. 3. System architecture overview

During the second stage in our framework, a augmented prefix tree acceptor APTA(S^+) with the input of message format sequence is constructed. The third stage is the state labeling for APTA, which is aiming at classifying different massage domain type for restricting over-merges. At last stage we derive the finite state machine with the modified RNPI algorithm thereby providing significant complexity reduction.

3.3 Message Format Selection

The APTA is used as a starting point to construct the protocol state machine. This is done by finding the minimal DFA that is consistent with the APTA. To find such a DFA, we can leverage existing algorithms (i.e. RPNI) that start from APTA and successively merge pairs of states. Clearly, states with different labels can never be merged. We introduce an algorithm that assigns different labels to the states of APTA. This restricts the possible merges, since only states with the same label may be merged. Finally, we use the modified RPNI to obtain a minimal DFA that is consistent with that labeling. This minimal DFA represents our state machine[8].

The goal of the state labeling algorithm is to find states in the APTA that are different. By assigning different labels to these states, we can prevent them from being merged. In this work, we leverage the following heuristics about network protocol to accomplish the state labeling.

Heuristics 1: As we observed, frequent interactive handshakes are usually performed at the beginning of a session for both server-client and P2P models. For the most representative interactive modes, the basic steps that an application session must take. As the key interactive initialization steps generally occur at the first few packets of a session. As an example, in the FTP command and control protocol, a login is required before other commands become available. In POP3, a “PASS” operation must be performed to connect to a share before file operations can be issued. In addition, certain commands may lead the server away from a state where it can perform these actions. For instance, a QUIT command in FTP make previously available commands impossible to execute.

Heuristic 2: Our state labeling algorithm attempts to identify states that represent similar application conditions. That is, we attempt to identify cases in which an application can process similar commands, based on the sequence of messages that it previously received. To this end, we regard the similar message as same type and to assign same labels so that we can extract simple patterns from the observed application sessions.

Heuristic 3: A prerequisite requires that, for the server to be in a state where it can meaningfully process a message of type m, it must first receive a message of type r. The message of type r is a message that always occurs before m. That is, in all application sessions, a message of type r was found before m. A common pattern in application layer network protocols is that a message or a sequence of messages must be sent before the server can perform certain actions. This is to capture the case where a connection or login message must be sent before message m.

Heuristic 4: In addition to the techniques described above, we also use a simple heuristic to detect end-states in the protocol state machine. It is common for application layer protocols to have one (or more) message types that request the termination of the protocol session. To detect those message types, we simply look for messages that, throughout all observed application sessions, appear only last in a session.

3.4 The Improvement of RPNI Algorithm

Clearly, in this work, states with different labels can never be merged. However, in our training set, all states of APTA are labeled accept. Thus, the result of directly applying an existing algorithm would be an over-general DFA with only a single state. To address this problem, we introduce an algorithm that assigns different labels to the states of APTA as discussed above. To this end, we embed the state merges algorithm into the original RPNI and so that to realize the assumed goal. The improvement of RPNI algorithm denoted as pseudo-code is demonstrated as follow.

```
input $S^+$, $S^-$
output A DFA consistent with $S^+$, $S^-$
begin
A←APTA($S^+$)        // N denotes the number of states of APTA(S+)
A←LABEL(A)                // label the APTA
$\pi$←{{0}, {1}, . . . , {N − 1}}      // One state for each prefix
                                              //according to standard order <
for i = 1 to |$\pi$|-1              // Loop over partition subsets$\pi$
   for j = 0 to i-1                 // Loop over subsets of lower rank
      $\pi'$←$\pi\backslash\{B_j,B_i\} \cup \{B_i \cup B_j\}$          // Merging $B_i$ and $B_j$
      $A/\pi'$ ← derive (A, $\pi'$)
      $\pi''$←determ_merging ($A/\pi'$)
      if compatible ($A/\pi''$,$S^-$) then        // Deterministic parsing of $S^-$
         $\pi$←$\pi''$
         break                     // Break j loop
      end if
   end for                  // End j loop
end for                 // End i loop
return $A/\pi$
```

4 Evaluation

In this section, we present the experimental evaluation of our approach. We have evaluated our method through implementing 3 different protocols (HTTP, SMTP, FTP) as shown in TABLE 1. We present the experimental evaluation of our approach. We perform an analytical comparison of our proposed algorithms with the original RPNI. Moreover, we demonstrate the derived DFA applying our modifying RPNI algorithm, the quality of specification produced by our method is also evaluated.

Table 1. Example of session summary for derived DFA

No.	Derived DFA	Positive Session(#)	Negative Session(#)
1	A_{HTTP}	HTTP(8762)	FTP(1423),SMPT(1356),POP3(1209)
2	A_{SMTP}	SMTP(2456)	HTTP(2134),POP3(1432),FTP(2134)
3	A_{FTP}	FTP(6589)	HTTP(2058),SMTP(2317),POP3(1653)

In our training set, all states of APTA are labeled accept. Thus, the result of directly applying an existing RPNI algorithm would be an over-general DFA with only a single state. Otherwise, even we learn the finite state machine adopting the original RPNI algorithm with corresponding negative sample, the result also is an over-general DFA. The following test validates the conclusion. In the experiment the positive examples are the network application with HTTP protocol, and the negative examples are the network application with FTP protocol. We learned the DFA with the original RNPI algorithm, the derived DFA is showed as Fig. 4. We find that the extracted DFA is exactly an over-general DFA. The result indicated the essential difference between HTTP and FTP protocol. Clearly, the DFA cannot describe the protocol structure of HTTP at all.

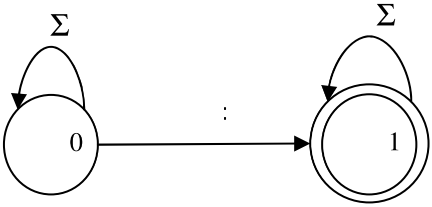
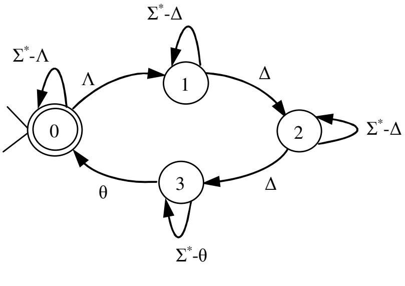


Fig. 4. An example of derived DFA with original RNPI

The derived DFA applying our method are demonstrated as Fig. 5–Fig.7 .

In the following section, we validate the soundness of derived DFA. We analyze the real-world network application implementing the protocol of HTTP, SMTP and FTP. The network traces used here are not the ones depicted in TABLE 1. The training network traces applied here are showed as TABLE 1. These traces regarded as input are analyzed by the derived DFA, the ratio rate of getting to the accepting state of derived DFA is showed as TABLE 2.



Λ={GET, HEAD, POST, PUT, OPTIONS }, Δ={1.1, 1.0, 0.9}
θ={100, 101, 200, 201, 202, 203, 204, 205, 206, 300, 301, 302, 303, 304, 305, 400, 401, 403, 404, 405, 406, 408, 411, 412, 414, 415, 505}

Fig. 5. Derived DFA for HTTP

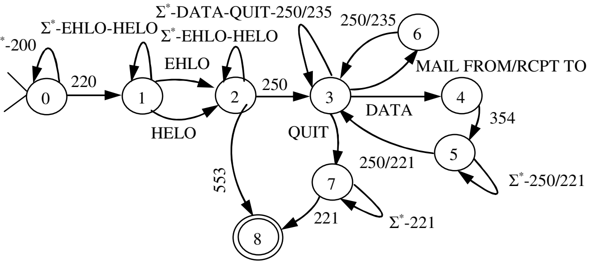
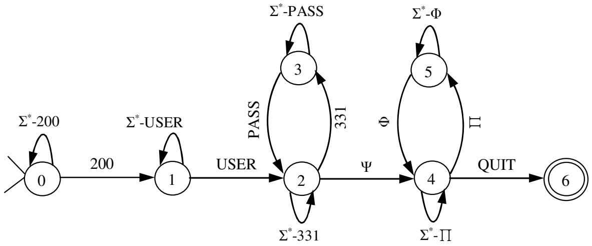


Fig. 6. Derived DFA for SMTP



Ψ={150,227,230,213,227,200,250,421,530,},∏={TYPE,PASV,SIZE,RETR,CWD,PORT,RES
T,LIST,ABOR,MODE,STOR,RMD,MKD,STAT,
NOOP},Φ={125,150,200,213,221,226,227,250,350,352,421,425,426,450,500,550,530}

Fig. 7. Derived DFA for FTP

Table 2. Test of soundness for derived DFA

NO.	Derived DFA	Type	Session example(#)	Session analyzed successfully(#)	Ratio
1	A_{HTTP}	HTTP	2423	2423	1
2	A_{SMTP}	SMTP	321	305	0.95
3	A_{FTP}	FTP	2100	1869	0.89

Out of 4844 total sessions, we were able to parse 4597 sessions (94.9%) successfully, which show that the ratio of successful analysis is very high. We check that the remaining 247 sessions (5.1%) were all using encryption, which we cannot handle properly as one of the limitations of our system is its inability to handle encrypted traffic. This shows that our system can fully parse (unencrypted) real-world traffic, generated by a number of clients and sent to a different mail server implementation than the one used to infer the protocol specifications.

5 Conclusion

In this work, we applied the improvement of RPNI algorithm to model protocol specification as DFA from the network traces input. Our experiments with real-world protocols and server applications demonstrate that the derived DFA take on well differentiating capability for different types of application protocols. We believe that the techniques that we introduce in this paper will be useful for related security policy.

References

1. Hingston, P.: A genetic algorithm for regular inference. In: Proc. Genetic and Evolutionary Computation Conf., San Francisco, USA, pp. 1299–1306 (2001)
2. Chomsky, N., Miller, G.A.: Pattern conception. Rapport technique. AFCRC-TN-5757 (ASTIA Document AD 110076) (1957)
3. Chodorowski, J., Miclet, L.: Applying grammatical inference in learning a language model for oral dialogue. In: Honavar, V.G., Slutzki, G. (eds.) ICGI 1998. LNCS (LNAI), vol. 1433, p. 102. Springer, Heidelberg (1998)
4. Oncina, J., Garcia, P.: Inferring regular languages in polynomial update time. Pattern recognition and image analysis: selected papers from the IVth Spanish Symposium, Granada, Spain, 49–61 (1990)
5. Parekh, R., Honavar, V.: Learning DFA from simple examples. In: Proc. Algorithmic Learning Theory, 8th International Workshop, Sendai, Japan, pp. 116–131 (1997)
6. Yin, H., Song, D., Manuel, E., Kruegel, C., Kirda, E.: Panorama: Capturing System-Wide Information Flow for Malware Detection and Analysis. In: ACM Conference on Computer and Communications Security, Alexandria, VA (October 2007)
7. Caballero, J., Song, D.: Polyglot: Automatic Extraction of Protocol Format using Dynamic Binary Analysis. In: ICCS 2007, pp. 317–329 (October 2007)
8. Comparetti, P.M., Wondracek, G., Kruegel, C., Kirda, E.: Prospex: Protocol Specification Extraction. In: IEEE Symposium on Security and Privacy, IEEE Computer Society Press, USA (May 2009)

Gene-Certificate Based Model for User Authentication and Access Control

Feixian Sun[1,2]

[1] School of Computer Science, Zhongyuan University of Technology
451191 Zhengzhou, China
[2] Zhengzhou Key Lab of Network Security Assessment
450007 Zhengzhou, China
sjysfx781@163.com

Abstract. Inspired by the principles of the human natural trust, a gene-certificate based model for user authentication and access control is proposed in this paper. With the formal definitions of gene-certificate, network-family, family member and gene defined, the algorithms of gene assignment, access control policy, gene signature, and gene-certificate generation, are described. Following that, the methods of network family construction and gene-certificate based authentication and access control, are designed. Stimulation results and theoretical analysis show that the presented model is valid, and it has the features of better safety. Thus, it provides an effective novel solution to network security.

Keywords: gene certificate, user authentication, access control.

1 Introduction

Traditionally, authentication and access control are implemented in different modules. So the access control problems, which result from the penetration of conventional authentication mechanisms, exist obviously [1]. On the other hand, the PKI based user authentication methods have many deficiencies, such as ambiguity of subject information, complicated identification process, high communication traffic, and so on [2]. An Integration solution, which provides both user authentication and access control in a single module to avoid any possible security breach between these two protecting mechanisms, has been firstly proposed by Harn et al. [3]. However, this method is based on static plaintext password, the security is still poor. Jan et al [4] proposed two integrated schemes for user authentication and access control, which provide an efficient updating process for the modification of access rights and allow servers to simplify verification processes for multiple access requests of a user, but these two schemes are very weak to the impersonation attack. The integrated user authentication and access control schemes without public key cryptography or using smart cards have also been presented [5, 6], they have the merits of no user-sensitive data stored on servers, no storage for access list or capability list, low computational cost, freedom of choosing users' passwords, and mutual authentication, but these two methods are difficult to synchronize the time between clients and servers.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 228–235, 2010.

There is a analogy of the trust method between computer n and Human Immune System (HIS). They both have to differentiate between non-self and self, and keep stability in a changing environment [7]. Due to desirable characteristics, such as diversity, immune-memory, distributed, self-learning, and self-adaptation, HIS has attracted researchers' attentions, and exciting results have been obtained [8-14].

Inspired by principles of the human natural trust, a Gene-certificate Based model for User Authentication and Access Control (GBUAAC) is proposed. The remaining of the paper is organized as follows. In Section 2, the theoretical model for user authentication and access control is proposed. In Section 3, stimulation and theoretical analysis are provided. Finally, Sections 4 contains our conclusion and future work.

2 Proposed Theoretical Model

In GBUAAC, a network is regarded as a *network-family*, a network user is taken as a *visitor*, network services are looked as *children* (a *child* is not always a real computer, and a computer can be presented with many children according to the services it provides), and their definitions are formally describes as follows.

2.1 Formal Definitions

In our proposed model, each network-family member is signed and issued a unique gene certificate. Given C denotes the set of gene certificates and it is defined by:

$$C = \left\{ \begin{array}{l} \left\langle \begin{array}{l} ver, name, type, desc, family\_gene, ext, \\ member\_gene, rule, sign\_id, sign\_val \end{array} \right\rangle \mid ver, name, desc, rule \\ sign\_id, family\_gene, ext, member\_gene, sign\_val \in N, type \in T \end{array} \right\}. \quad (1)$$

where $N = \bigcup_{i=1}^{\infty} \{a, \cdots, z, ., \_, 0, \cdots, 9, \text{空格}\}^i$, and $T = \{parent, child, visitor\}$.

In GBUAAC, the *network-family* (*F*) is composed of parents, children and visitors, and its forma*l* definition is given as follows.

$$F = \left\{ \begin{array}{l} m \mid m = \left\langle gene\_certificate, inheritance\_password \right\rangle, \\ gene\_certificate \in C, inheritance\_password \in N \end{array} \right\}. \quad (2)$$

where the *inheritance_password* of the *parent* family member is used to sign gene-certificates for its *children*, and the *inheritance_password* of *children* and *visitors* is used for encryption.

Judging by equation (1), we know the family member of GBUAAC includes three types: *parent*, *child* and *visitor*. For $\forall m \in F$, the gene of m ($gene_m$) is defined by:

$$\begin{array}{l} gene_m = m.gene\_certificate.family\_gene \parallel \\ m.gene\_certificate.member\_gene. \end{array} \quad (3)$$

where the *family gene* is used to differentiate the sub networks, and the *member* gene is used to distinguish the family members within a sub network.

We define *parent* ($p, p \in F$) of GBUAAC is a special family member, which can own *children*, *visitors*, and child *parents*. According to equation (3), we have:

$$\begin{aligned} gene_p = p.gene\_certificate.family\_gene \,\| \\ p.gene\_certificate.member\_gene. \end{aligned} \tag{4}$$

Let $p(p \in F)$ denote the parent of $m(m \in F)$, and $\langle k_{pub}, k_{pri} \rangle$ is a public/private key pair. According to the inheritance attribute of computer networks, we define:

$$m.gene\_certificate.family\_gene = gene_p \tag{5}$$

$$m.gene\_certificate.member\_gene = k_{pub} \tag{6}$$

$$m.inheri\,tan\,ce\_password = k_{pri} \tag{7}$$

We notice that the *network-primogenitor* ($p_0, p_0 \in F$) is a special parent, which has no parent. So we have:

$$p_0.gene\_certificate.family\_gene = null \tag{8}$$

In real network environments, the *family_gene* of each family member within a same sub network can be realized by the net-id, and the *member_gene* can be implemented by the physical address (MAC).

2.2 Gene Assignment

In GBUAAC, the gene assignment is to assign the *family_gene*, *member_gene*, and *inheritance_password* for the family members, this procedure is realized by *parent*s, and the Gene Assignment Algorithm (GAA) is formally described as follows.

Algorithm 1. GAA

```
Input: m,p;  //p is the parent of m
Begin
 assign family_gene for m;  //equation (5)
 generate a public/private key pair $\langle k_{pub}, k_{pri} \rangle$;
 assign member_gene for m;  //equation (6)
 assign inheritance_password for m;  //equation (7)
 generate $gene_m$;  //equation (3)
End.
```

We notice that the different family members within a subnet-family can't have the same *member_gene*.

2.3 Access Control Policy

Within GBUAAC, the gene certificate field *rule* of the family member *m* represents the access control policies, which authorizes *m* to perform a set of actions on the set of network-family resources.

In our proposed model, the constitution of access control policy of each visitor (network users) is based on the security policies of the whole network, and the constitution method is described as follows.

Step 1: constitute *home access rights*. Let L_m^h denote the *home access rights* list of m, L_m^h includes the members, which are in the same subnet-family with m, and they can be accessed by m.

Step 2: constitute *family access rights*. Let p represent the *parent* of m, L_m^f denote the *family access rights* list of m, and it is inherited from p (by this means, m can access its elder family members).

Step 3: constitute *access denied list*. Let L_m^u denote the *access denied list* list of m, L_m^u includes all the family members which can not be accessed by m.

Step 4: generate access control policy. $m.gene\_certificate.rule = L_m^h + L_m^f + L_m^u$.

2.4 Gene Signature

For each $m \in F$, let p represent the parent of m, H denote the hash function (MD5 or SHA-1), $E_k(s)$ represents encrypting s with key k. We let:

$$\begin{aligned} M = {} & m.gene\_certificate.ver \,||\, m.gene\_certificate.name \,||\\ & m.gene\_certificate.type \,||\, m.gene\_certificate.desc \,||\\ & m.gene\_certificate.family\_gene \,||\, m.gene\_certificate.ext \,||\\ & m.gene\_certificate.member\_gene \,||\, m.gene\_certificate.rule \,||\\ & m.gene\_certificate.sign\_id \end{aligned} \tag{9}$$

Let $h = H(M)$, $m.gene\_certificate.sign\_val = E_{p.inheri\tan ce\_password}$.

2.5 Generate Gene_Certificate

After gene assignment, access control policy constitution and gene signature, the gene-certificate of m can be generated, and the algorithm is described as follows.

Algorithm 2. GGA //Gene_certificate Generation Algorithm

```
Input: m,p;  //p is the parent of m
Begin
  Set the gene_certificate version for m;
  Set the gene_certificate name for m;
  Set the gene_certificate type for m;
  Set the gene_certificate description for m;
  Assign gene for m;   //algorithm 1. GAA
  Set the gene_certificate extension information for m;
  Set the access control policy for m; //According to 2.3
  gene signature for m;  //According to Segment 2.4
End.
```

2.6 Construct Network-Family Evolvement Map

According to the definitions of network family and family members, a network is considered a network-family, which is composed of sub network families. Sub network families, which denote sub networks, are evolved from the network primogenitor, and the network-family evolvement map is illustrated in Fig. 1.

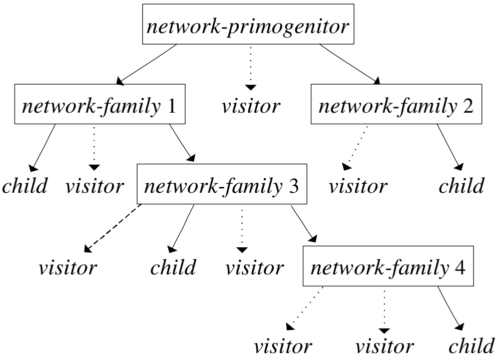


Fig. 1. Network-family evolvement map

2.6.1 Construct Network-Primogenitor

In GBUAAC, the *network-primogenitor* (p_0) is the ancestor. According to equation (8), the *family_gene* of the *network-primogenitor* is empty, and its gene certificate is self-signed. The Network Primogenitor Construction Algorithm (NPCA) is formally described as follows.

Algorithm 3. NPCA

```
Begin
  initialize network family evolvement map tree (T);
  New($p_0$);  //$p_0 \in$ F
  $p_0$.gene_certificate.name=primogenitor;
  $p_0$.gene_certificate.type=parent;
  $p_0$.gene_certificate.family_gene=null; //equation (8)
  generate a public/private key pair;
  $p_0.gene\_certificate.member\_gene = k_{pub}$;  //equation (6)
  $p_0.inheri\,tan\,ce\_password = k_{pri}$;  //equation (7)
  Initialize $p_0.gene\_certificate.rule$;
  Generate a self-signed gene certificate for $p_0$;
  Insert ($p_0$, T);
End.
```

2.6.2 Generate Family Members

In GBUAAC, managers, users, and services are converted into *parent*s, *visitor*s, and *children*. Through gene assignment, access control policy constitution, and gene signature, the accepted m is assigned a gene certificate, then m is inserted into T as a node of p, and the algorithm is described as follows.

Algorithm 4. FMGA //Family Member Generation Algorithm

```
Begin
 m apply to its parent p for registration;
   If (the application is passed)
    {create a new node n, n•F;
     Call GAA;  //Algorithm 1. GAA
  Set the access control policy for m; //see segment 2.3
  Call GGA; //Algorithm 2. GGA
  insert n into T as a node of p;
  }
End.
```

2.7 User Authentication and Access Control

Fig. 2 illustrates the gene-certificate based algorithm for user authentication and access control.

3 Stimulation and Theoretical Analysis

According to Fig. 1, we built the evolvement-map of a real network that was composed of four subnets. To draw a convictive conclusion, we compared the stimulation results with the method proposed by Eigeles [15] under the same circumstance: CPU: Intel Celeron 2.4G, Memory: 256M, OS: Red Hat Linux 7.2, Program Language: C.

The public/private key pairs were generated with RSA, and the key size was 1024 bits. The message-digest and digital-signature algorithms were SHA-1 and RSA, respectively. The experiments were evaluated by the speed of certificate generation and authentication. Fig. 3 shows the comparison of certificates generation speed between the gene certificates and X.509 (v3) certificates. Fig. 4 illustrates the comparison of the authentication speed between GBUAAC and I3A [15]. In these two figures, the value of each group is the average time of 30 times experiment.

From Fig. 3 and Fig. 4 we see that, the speed of certificates generation and identity authentication of GBUAAC is faster than the existing models or methods. The good results of the performance test of GBUAAC root in the following facts. Firstly, the gene certificate format of GBUAAC is simpler than that of X.509. Secondly, the times of verifying certificate signature in GBUAAC is 1 time less than I3A. Lastly, being without certificate revocation list in GBUAAC, the gene certificate management system becomes very simple.

Firstly, there isn't any password to transmit in the proposed model, so the network attacks, such as password guessing, and network monitoring are invalid to GBUAAC. Secondly, comparing with PKI, there isn't RA in GBUAAC, hence the communication traffic decreases greatly. Finally, the detailed identity information of each family-member, such as the name, type, derivation, origin, and etc., can be confirmed exactly through analyzing its family gene and member gene. Therefore, the ambiguity of subject information of the gene certificate does not exist in GBUAAC.

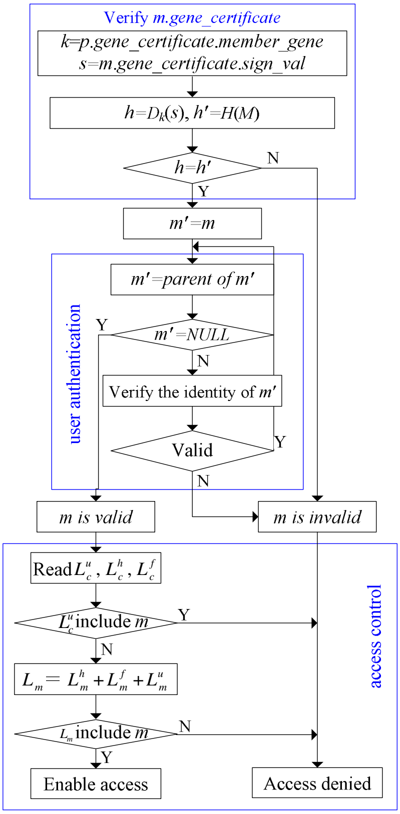


Fig. 2. Gene_certificate based algorithm for user authentication and access control

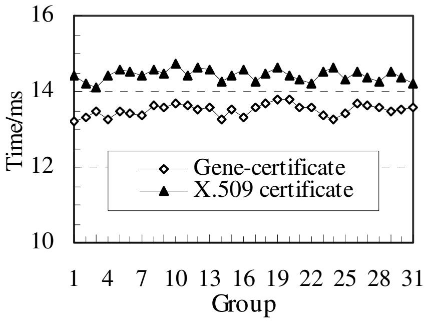


Fig. 3. Comparison of certificate generation speed

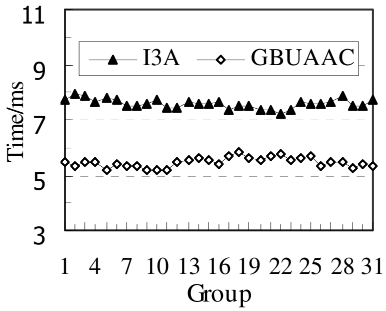


Fig. 4. Comparison of authentication speed

4 Conclusion

This paper abstracted and extended the concept of gene from biological immune system, and proposed a gene-certificate based model for user authentication and access control. The presented model extracts the virtues of PKI/PMI and the natural trust method of mankind, it can ensure the family-members pure origin by family-gene and member-gene, and it can avoid the possible security breach between user authentication and access control by integrating these two mechanisms within a single module. Not only the proposed model is feasible, but also it has the features of better safety. Thus, it provides an effective novel solution to network security.

References

1. Balthrop, J., Forrest, S., Newman, M.E.J., Williamson, M.M.: Technological Networks and the Spread of Computer Virus. Science 304, 527–529 (2004)
2. Ellison, C., Schneier, B.: Ten Risks of PKI: What You're Not Being Told About Public Key Infrastructure. Computer Security J. 16, 1–7 (2000)
3. Harn, L., Lin, H.Y.: Integration of User Authentication and Access Control. IEEE Proceedings - Computers and Digital Techniques 139, 139–143 (1992)
4. Jan, J.K., Tseng, Y.M.: Two Integrated Schemes of User Authentication and Access Control in a Distributed Computer Network. IEEE Proceedings - Computers and Digital Techniques 145, 419–424 (1998)
5. Chien, H.Y., Jan, J.K.: An Integrated User Authentication And Access Control Scheme Without Public Key Cryptography. In: 2003 IEEE International Conference on Security Technology, pp. 137–143. IEEE, New York (2003)
6. Chen, Y.C., Yeh, L.Y.: An Efficient Authentication and Access Control Scheme Using Smart Cards. In: 11th International Conference on Parallel and Distributed Systems, pp. 78–82. IEEE Computer Soc., Los Alamitos (2005)
7. Li, T.: Computer Immunology. Publishing House of Electronics Industry, Beijing (2004)
8. Sun, F., Kong, M., Wang, J.: An Immune Danger Theory Inspired Model for Network Security Threat Awareness. In: 2010 International Conference on Multimedia and Information Technology, vol. 2, pp. 93–95 (2010)
9. Sun, F., Han, X., Wang, J.: An Immune Danger Theory Inspired Model for Network Security Monitoring. In: 2010 International Conference on Challenges in Environmental Science and Computer Engineering, vol. 2, pp. 33–35 (2010)
10. Sun, F., Wu, Z.: A New Risk Assessment Model for E-Government Network Security Based on Antibody Concentration. In: 2009 International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, pp. 119–121 (2009)
11. Sun, F., Xu, F.: Antibody Concentration Based Method for Network Security Situation Awareness. In: 3nd International Conference on Bioinformatics and Biomedical Engineering, vol. 1, pp. 1–4 (2009)
12. Sun, F., Zheng, Q., Li, T.: Immunity-Based Dynamic Anomaly Detection Method. In: 2nd International Conference on Bioinformatics and Biomedical Engineering, vol. 1, pp. 644–647 (2008)
13. Liu, F., Pan, X.Y.: Block Motion Estimation Based on Immune Clonal Selection. J. of Software 18, 850–860 (2007)
14. Kim, M., Seo, J.: Network Anomaly Behavior Detection Using an Adaptive Multiplex Detector. In: Gavrilova, M.L., Gervasi, O., Kumar, V., Tan, C.J.K., Taniar, D., Laganá, A., Mun, Y., Choo, H. (eds.) ICCSA 2006. LNCS, vol. 3982, pp. 154–162. Springer, Heidelberg (2006)
15. Eigeles, D.: Intelligent Authentication, Authorization, and Administration (I3A). Information Management & Computer Security 14, 5–23 (2006)

A New Data Integrity Verification Mechanism for SaaS[*]

Yuliang Shi, Kun Zhang, and Qingzhong Li

School of Computer Science and Technology, Shandong University,
Jinan, P.R. China
liangyus@sdu.edu.cn, jackie_119@mail.sdu.edu.cn,
lqz@sdu.edu.cn

Abstract. Recently, the Software-as–a-Service (SaaS) model has been gaining more and more attention. In SaaS model, both applications and databases will be deployed in servers managed by untrustworthy service providers. Thus, the service providers might maliciously delete, modify or falsify tenants' data due to some reasons, which brings great challenge to adoption of SaaS model. So this paper defines data integrity concept for SaaS data storage security, which could be measured in terms of durable integrity, correct integrity and provenance integrity. Basing on the meta-data driven data storage model and data chunking technology, SaaS data integrity issues will be mapped as a series of integrity issues of data chunks. Via cyclic group, data chunks traversal approach for verification is presented, and then SaaS data integrity verification can be realized based on the integrity verification of data chunks. Also, we demonstrate the correctness of the mechanism through analysis in this paper.

Keywords: SaaS, Integrity Verification, Data Chunk, Cyclic Group.

1 Introduction

Software-as-a-Service, i.e. SaaS, is a new software delivery model with the development of network and maturity of application software. In SaaS model, applications and databases are hosted at the untrustworthy service provider's servers. Since service provider may delete, modify, falsify and fabricate tenant's data for some commercial reasons, secure data storage becomes the biggest challenge in SaaS.

Example 1: In a CRM system of SaaS model, customers' information is stored at the service provider's servers. However, service providers would delete some data which is seldom accessed, so they can reclaim the storage space for other tenants. And service provider would collude with some competitors to modify or falsify some customer information for economic interest.

This paper focuses on the case that service providers are not trustworthy. For secure data storage in SaaS model, we firstly define the concept of data integrity in SaaS model, which includes durable integrity, correct integrity and provenance

[*] The research is supported by the National Natural Science Foundation of China under Grant No.60673130 and No.90818001, the Natural Science Foundation of Shandong Province of China under Grant No.Y2007G24 and No.Y2007G38.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 236–243, 2010.

integrity. Then based on a meta-data driven multi-tenancy data-sharing storage model, all the storage security issues about tenants' data will be mapped as a series of integrity issues about data chunks by data chunking technology. Through the integrity verification of data chunks, the secure storage of tenants' data is assured.

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 gives the definition of data integrity in SaaS model, presents data chunk based meta-data driven multi-tenancy data-sharing storage model and analyzes the association between data integrity issue from tenants' logical view and a series of data chunks integrity issues from physical view. Section 4 presents the data integrity verification approach and makes an analysis. Finally section 5 concludes the paper.

2 Related Works

For secure data storage in cloud computing, [1] proposed an effective and flexible distributed schema by utilizing the homomorphic token with distributed verification of erasure-coded data. This schema achieved the integration of storage correctness insurance and data error localization and supported efficient dynamic operations on data blocks. Reference [2] assigned the task of ensuring the integrity of dynamic data storage in cloud computing to a third party auditor, which eliminated the involvement of client. Reference [3] introduced HAIL (High-Availability and Integrity Layer), which is a distributed cryptographic system that permits a set of servers to prove to a client that a stored file is intact and retrievable. HAIL extends the basic principles of RAID into the adversarial setting of the cloud and is a remotely file integrity checking protocol. However, these approaches just consider the data integrity in the database layer, which is not inefficient for SaaS model where software and databases are both deployed at service provider's platform.

For secure data storage at untrusted host, [4] gave a survey of data integrity and data completeness approach in Database-as-a-Service model. These methods focus on the signature-based, challenge-response methods and probability-based approaches. However, signature-based approaches are based on the ordered data, challenge-response methods are used to adopt the DBMS kernel. Reference [5] inserted certain fake tuples into the real data and verified query integrity by checking the fake tuple in the result. Reference [6] presented the dual encryption approach, where certain data are encrypted with different keys and query integrity could be checking by "cross examination". Reference [7] proposed PORs model, which enables an archive or back-up service (prover) to produce a concise proof that a user (verifier) can retrieve a target file F. Reference [8] introduced a model for provable data possession (PDP) that allows a client that has stored data at untrusted server to verify that the server possesses the original data without retrieving it, which utilize public key based homomorphic tags for auditing the data file, thus providing public verifiability. However, all these methods are not available directly for SaaS model.

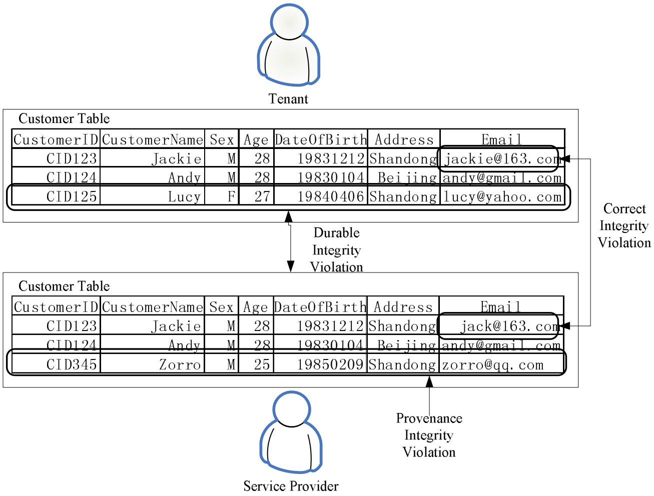
Reference [9] presented a data privacy preserving mechanism based on tenant customization for SaaS. In this mechanism, [9] introduced a multi-tenancy data-sharing storage model, which is used for privacy preservation based on the data fragmentation by introducing of FragID. By fragmentation, it is difficult or impossible for service providers to get a complete data record caused by the hidden of association between data chunks. This feature could be used in our paper.

3 Data Integrity for SaaS

For data storage security issue, we first define the concept of data integrity in SaaS, and then present the data chunks based meta-data driven multi-tenancy data-sharing storage model, which is the basis of data integrity for SaaS. Then we analyze the association between data integrity issue in tenants' logical view and data chunks integrity issue in physical storage.

3.1 Data Integrity for SaaS

In SaaS model, data integrity ensures secure storage of tenants' data, which we use data integrity to protect tenants' data from being deleted, modified and falsified.



Customer Table

CustomerID	CustomerName	Sex	Age	DateOfBirth	Address	Email
CID123	Jackie	M	28	19831212	Shandong	jackie@163.com
CID124	Andy	M	28	19830104	Beijing	andy@gmail.com
CID125	Lucy	F	27	19840406	Shandong	lucy@yahoo.com

Customer Table

CustomerID	CustomerName	Sex	Age	DateOfBirth	Address	Email
CID123	Jackie	M	28	19831212	Shandong	jack@163.com
CID124	Andy	M	28	19830104	Beijing	andy@gmail.com
CID345	Zorro	M	25	19850209	Shandong	zorro@qq.com

Fig. 1. A Simple example of data integrity in SaaS model

Definition 1: Data Integrity in SaaS model: Given a relation R (A_1, $A_2 \ldots A_n$) from tenants' view, a certain data tuple is denoted by tuplei (a_{i1}, $a_{i2} \ldots a_{in}$). The tuple is in state of integrity if and only if the three following conditions are satisfied:

(1) Durable integrity: data is retrievable, i.e. tenants' data could not be deleted without awareness of tenants who own these data.

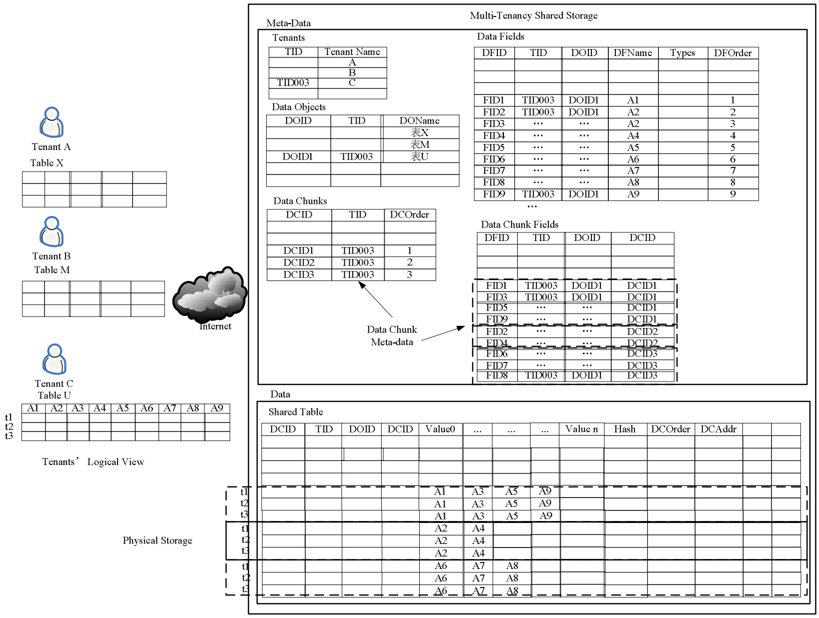
(2) Correct integrity: data is correct, i.e. only authorized user could update the data stored in the service provider's servers.

(3) Provenance integrity: every data item appeared in the tenants' logical view has its legitimate source, i.e. data should not be falsified by service provider.

Example 2: Given a simple customer table in CRM system in Example 1, as shown in Fig. 1. Service providers may do harm to tenants' data integrity. For example, service provider may delete the record whose Customer ID is CID125, which violates durable integrity. Correct integrity may be violated by changing email of customer. Service providers may add some non-existence customer information to tenants' database hosted by service providers, which violates provenance integrity.

3.2 Data Chunk Based Meta-data Driven Data Storage Model

Based on the previous work [9] on data privacy preservation in SaaS, we use the data chunks based meta-data driven data storage model as the basis of data integrity verification mechanism for SaaS. In this paper, we combine the data privacy protection and data integrity verification based on this data shared storage model, as shown in Fig. 2.



Tenants

TID	Tenant Name
	A
	B
TID003	C

Data Objects

DOID	TID	DOName
		表X
		表M
DOID1	TID003	表U

Data Chunks

DCID	TID	DCOrder
DCID1	TID003	1
DCID2	TID003	2
DCID3	TID003	3

Data Fields

DFID	TID	DOID	DFName	Types	DFOrder
FID1	TID003	DOID1	A1		1
FID2	TID003	DOID1	A2		2
FID3	···	···	A2		3
FID4	···	···	A4		4
FID5	···	···	A5		5
FID6	···	···	A6		6
FID7	···	···	A7		7
FID8	···	···	A8		8
FID9	TID003	DOID1	A9		9

···

Data Chunk Fields

DFID	TID	DOID	DCID
FID1	TID003	DOID1	DCID1
FID3	TID003	DOID1	DCID1
FID5	···	···	DCID1
FID9	···	···	DCID1
FID2	···	···	DCID2
FID4	···	···	DCID2
FID6	···	···	DCID3
FID7	···	···	DCID3
FID8	TID003	DOID1	DCID3

Table U

	A1	A2	A3	A4	A5	A6	A7	A8	A9
t1									
t2									
t3									

Shared Table

	DCID	TID	DOID	DCID	Value0	...	...	...	Value n	Hash	DCOrder	DCAddr		
t1					A1	A3	A5	A9						
t2					A1	A3	A5	A9						
t3					A1	A3	A5	A9						
t1					A2	A4								
t2					A2	A4								
t3					A2	A4								
t1					A6	A7	A8							
t2					A6	A7	A8							
t3					A6	A7	A8							

Fig. 2. Data chunk based meta-data driven data storage model

Meta-data tables include tenants table, data objects table, data fields table, data chunks table and data chunk fields table. Tenants table describes the information of tenants, including tenant ID and tenant's name. Data object table describes the information about data table of tenants, including tenant object ID, tenant object name and the tenant ID. Data field table describe the information about the data field in each data objects, including data type, data order in the logical view of tenants, data object ID, tenant ID and other information. Data chunks table describe the data chunks, including data chunk ID, tenant ID and other information. Data chunk fields table describe the detailed information about each data chunk, including data field ID in the data chunk. Shared tables are prepared for tenants' business data. Service provider stores tenants' data based on the meta-data definition. The fields in each row of shared tables include tenant ID, data object ID, data chunk ID, data values and other customizable data fields, which could be used for data privacy and integrity verification.

As shown in Fig. 2, table U in tenant C's logical view are mapped into three data chunks in shared data tables according to $\{A_1, A_3, A_5, A_9\}$, $\{A_2, A_4\}$ and $\{A_6, A_7, A_8\}$ data chunking model.

3.3 Data Integrity Based on Data Chunks

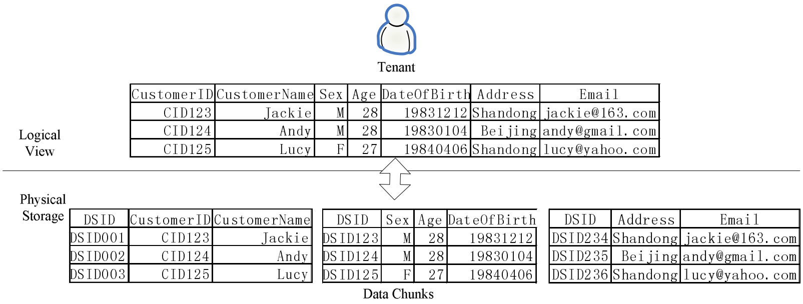
For secure data storage issue in SaaS model, data integrity issue of tenants' data could be mapped from logical view to a series of data integrity issues of data chunks from the physical storage. The association between data chunks of the same data record is hidden from service provider, which has been protected in [9]. Based on protection of these associations, service provider could not delete, modify or falsify all data chunks of the same data record at the same time. Tenant could utilize any data chunk he gets to check the data integrity of data record.

A data chunk is in state of integrity if and only if all these three conditions are satisfied:

(1) Durable integrity: the data chunk exists in the physical storage. This integrity could be done by checking the existence of data chunk of the unique value.

(2) Correct integrity: the data chunk has not be modified or tampered by authorized users. This integrity could be done by verification of signature of data chunks.

(3) Provenance integrity: the data chunk is originated from tenant. This integrity could be done by checking the signature of data chunks.



CustomerID	CustomerName	Sex	Age	DateOfBirth	Address	Email
CID123	Jackie	M	28	19831212	Shandong	jackie@163. com
CID124	Andy	M	28	19830104	Beijing	andy@gmail. com
CID125	Lucy	F	27	19840406	Shandong	lucy@yahoo. com

DSID	CustomerID	CustomerName
DSID001	CID123	Jackie
DSID002	CID124	Andy
DSID003	CID125	Lucy

DSID	Sex	Age	DateOfBirth
DSID123	M	28	19831212
DSID124	M	28	19830104
DSID125	F	27	19840406

DSID	Address	Email
DSID234	Shandong	jackie@163. com
DSID235	Beijing	andy@gmail. com
DSID236	Shandong	lucy@yahoo. com

Fig. 3. The integrity mapping between logical view and physical storage

So based on the integrity of data chunks, the data integrity issue could be assured through the integrity verification of data chunks. The mapping between data integrity issue from tenants' logical view and data chunks integrity from physical storage is follows, as shown in Fig.3.

(1) A single data record from tenants' view is in state of integrity if and only if all data chunks from physical storage are in state of integrity.

(2) If any one of data chunks from physical storage is not in state of instate, data record from tenants' view is not in state of integrity.

4 Data Integrity Verification for SaaS

This section firstly gives the data chunks traversal approach, and then introduces and analyzes the integrity verification method.

4.1 Data Chunks Traversal

In order to location data chunks quickly, we use the cyclic group as tenants' private key. A group G is called cyclic if there exists an element *a* in G such that G = <*a*> = {a^n | n is an integer}. In the finite cyclic group ordered by n, all the elements are generated by the generator "*a*". Thus, the tenants can select the generator *a* for a finite cyclic group ordered k from the <Z_n, $+_n$> Group as the key to realize orientation for relative data segmentations immediately by combining the DSOrder of information integrality verifying and other information.

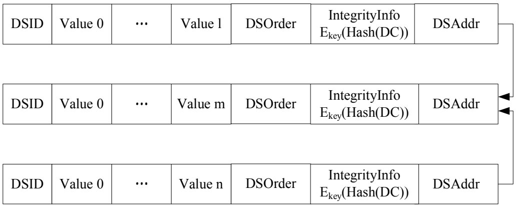


Fig. 4. Data chunk structure based on cyclic group

The basic data chunk structure based on cyclic group is shown in Fig.4. DSID is the unique identifier of data share, value$_i$ is data value, DSOrder denotes the order of the data share, DC is the combination of values of this data chunk, E_{key}(Hash(DC)) is the integrity information, which assures the correct integrity and provenance integrity. DSAddr is the data share address. The durable integrity is assured by the IntegrityInfo and DSAddr. The DSAddr is computed by the following equation (1).

$$DSAddr = E(ID \oplus a_{key}{}^{DSOrder}) \tag{1}$$

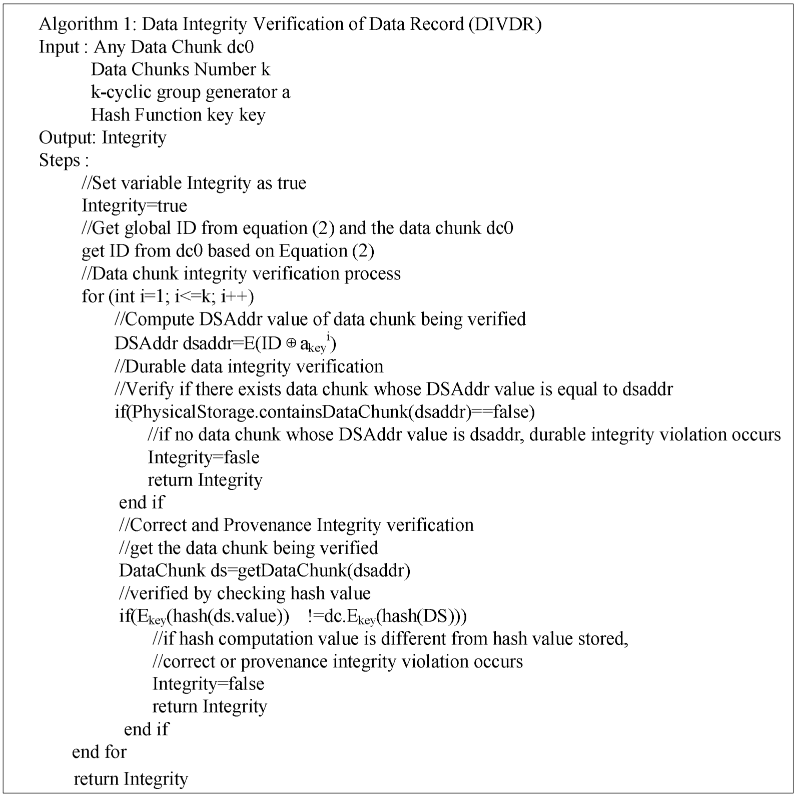
ID is the same identifier for the same record, key is the tenant's private key, a is the private k-cyclic group generator. Based on the IntegrityInfo, tenant could get the ID from equation (2). And then verify the durable integrity by checking the existence of data chunk.

$$ID = D_k(E_k(ID \oplus DSID)) \oplus DSID \tag{2}$$

4.2 Data Integrity Verification

Based on the mapping between data integrity from tenants' logical view and data chunks integrity from physical storage, tenants should verify the integrity of data chunks for securing data storage. Firstly, tenants should customize some integrity verification information. Then he could utilize any one of the data chunks he gets to traversal some or all data chunks for integrity verification, including durable integrity, correct integrity and provenance integrity. Based on the data chunks integrity verification, tenants assured the secure data storage. This paper focuses on the data integrity of single record.

The complete integrity verification method is shown in Fig. 5.



```
Algorithm 1: Data Integrity Verification of Data Record (DIVDR)
Input : Any Data Chunk dc0
        Data Chunks Number k
        k-cyclic group generator a
        Hash Function key key
Output: Integrity
Steps :
    //Set variable Integrity as true
    Integrity=true
    //Get global ID from equation (2) and the data chunk dc0
    get ID from dc0 based on Equation (2)
    //Data chunk integrity verification process
    for (int i=1; i<=k; i++)
        //Compute DSAddr value of data chunk being verified
        DSAddr dsaddr=$E(ID \oplus a_{key}^{i})$
        //Durable data integrity verification
        //Verify if there exists data chunk whose DSAddr value is equal to dsaddr
        if(PhysicalStorage.containsDataChunk(dsaddr)==false)
            //if no data chunk whose DSAddr value is dsaddr, durable integrity violation occurs
            Integrity=fasle
            return Integrity
        end if
        //Correct and Provenance Integrity verification
        //get the data chunk being verified
        DataChunk ds=getDataChunk(dsaddr)
        //verified by checking hash value
        if($E_{key}$(hash(ds.value))  !=dc.$E_{key}$(hash(DS)))
            //if hash computation value is different from hash value stored,
            //correct or provenance integrity violation occurs
            Integrity=false
            return Integrity
        end if
    end for
    return Integrity
```

Fig. 5. Algorithm for data integrity verification of data record

4.3 Analysis

Given a relation from tenant's logical view, it can be mapped to k data chunks from physical view, service providers attacked y data chunks ($1 \leqslant y \leqslant k$) and tenant take x ($1 \leqslant x \leqslant k$) of k to verify the integrity of data. When $x+y \leqslant k$, the verification probability of successfully checking the attack is

$$P(verify) = 1 - C_{k-y}^{x} \Big/ C_{k}^{x} \tag{3}$$

Analysis: When $x+y \leqslant k$, the count of picking y attacked data chunks from k is C_k^y. When there are no interaction between sets of verified data chunks and attacked ones, the verification process checks x data chunks from the remaining (k-y) ones. So the count is C_{k-y}^x. Since the total count of attacking y and verifying x is C_k^x/C_k^y, the probability of successfully verification is

$$P(verify) = 1 - C_{k}^{y} C_{k-y}^{x} \Big/ C_{k}^{x} C_{k}^{y} = 1 - C_{k-y}^{x} \Big/ C_{k}^{x} \tag{4}$$

5 Conclusions

In SaaS model, applications and databases are both hosted at the service providers' servers, secure data storage issue become the biggest challenge caused by the untrustworthiness of service providers. For secure data storage in SaaS model, we firstly define the concept the data integrity for SaaS, present meta-data driven multi-tenancy data-sharing storage model and map the data storage security issue to a series of data chunk integrity issues. So based on integrity verification for tenants' data chunks, storage security issue of tenants' data can be ensured as well.

However, we just only consider a simple situation in SaaS model. Based on these work, we would pay attention to the data set integrity verification, public verifiability and the data integrity recovery mechanism. Data set integrity verification would allow tenants to verify the integrity of any data record. The public verifiability would allow tenants to assign the verification task to a third party auditor without violating their privacy. Integrity recovery mechanism would make tenants' data back into state of integrity when attacks happened.

References

1. Wang, C., Wang, Q., Ren, K., Lou, W.: Ensuring data storage security in cloud computing. In: 17th IEEE International Workshop on Quality of Service (IWQoS 2009), pp. 1–9. IEEE Press, New York (2009)
2. Wang, Q., Wang, C., Li, J., Ren, K., Lou, W.: Enabling Public Verifiability and Data Dynamics for Storage Security in Cloud Computing. In: Backes, M., Ning, P. (eds.) ESORICS 2009. LNCS, vol. 5789, pp. 355–370. Springer, Heidelberg (2009)
3. Bowers, K., Juels, A., Oprea, A.: HAIL: a high-availability and integrity layer for cloud storage. In: Proceedings of the 2009 ACM Conference on Computer and Communications Security (CCS 2009), pp. 187–198. ACM, New York (2009)
4. Tian, X., Wang, X., Gao, M., Zhou, A.: Database as a Service – Security and Privacy Preserving. Journal of Software 21(5), 991–1006 (2010)
5. Xie, M., Wang, H., Yin, J., Meng, X.: Integrity Auditing of Outsourced Data. In: Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB 2007), pp. 782–793. ACM, New York (2007)
6. Wang, H., Yin, J., Perng, C., Yu, P.: Dual encryption for query integrity assurance. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), pp. 863–872. ACM, New York (2008)
7. Juels, A., Kaliski Jr., B.: PORs: proofs of retrievability for large files. In: Proceedings of the 2007 ACM Conference on Computer and Communications Security (CCS 2007), pp. 584–597. ACM, New York (2007)
8. Ateniese, G., Burns, R., Curtmola, R., Herring, J., Kissner, L., Peterson, Z., Song, D.: Provable data possession at untrusted stores. In: Proceedings of the 2007 ACM Conference on Computer and Communications Security (CSS 2007), pp. 598–609. ACM, New York (2007)
9. Zhang, K., Shi, Y., Li, Q., Bian, J.: Data Privacy Preserving Mechanism based on Tenant Customization for SaaS. In: Proceedings of 2009 International Conference on Multimedia Information Networking and Security (MINES 2009), pp. 599–603. IEEE Computer Society, Los Alamitos (2009)

An Exquisite Authentication Scheme with Key Agreement Preserving User Anonymity*

Mijin Kim, Seungjoo Kim, and Dongho Won**

School of Information and Communication Engineering,
Sungkyunkwan University, Suwon 440-746, Republic of Korea
{mjkim,skim,dhwon}@security.re.kr

Abstract. In 2009, Liao et al. proposed an exquisite mutual authentication scheme with key agreement using smart cards to access a network system legally and securely. Liao et al.'s scheme adopted a transformed identity (TID) to avoid identity duplication. However, we find out that an adversary may exploit TID to achieve offline guessing attack. Liao et al.'s scheme is also exposed to man-in-the-middle attack and their claimed theorems and proofs are incorrect. We conduct detailed analysis of flaws in the scheme and its security proof. This paper proposes an improved scheme to overcome these problems and preserve user anonymity that is an issue in e-commerce applications.

Keywords: Mutual authentication, key agreement, transformed identity, user anonymity.

1 Introduction

Authentication and key establishment are fundamental procedures to establish secure communications over public insecure networks. After Bellovin and Merritt proposed a Password-based Authenticated Key Exchange (PAKE) protocol secure against dictionary attacks in 1992 [1], many researchers proposed PAKE protocols [2,3,4,5] based on different cryptographic assumptions. Unlike previous key exchange schemes, PAKE protocols require an user to remember its short (human-memorizable) password and enable participating parties to share a common session key and authenticate each other. However, due to the short password length, special care must be taken when designing protocols to ensure that both the password and the key finally agreed remain secret.

Smart cards have been widely adopted in many cryptographic protocols due to their low cost, portability and cryptographic capabilities. PAKE also used

* This work was supported in part by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center (ITRC) support program supervised by the National IT Industry Promotion Agency (NIPA-2010-(C1090-1031-0005)) and in part by the Defense Acquisition Program Administration and Agency for Defense Development under Contract UD100002KD.

** Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 244–253, 2010.

a smart card as a security token for more efficient execution. However, the resources in smart cards are constrained; the computation and the communication overhead must be low for practical implementation.

In 2002, Hwang-Lee-Tang proposed a simple remote user authentication scheme [6]. Although the scheme can verify a legitimate user, the user and the server cannot achieve mutual authentication and session key agreement. Nor can the scheme avoid the time synchronization problem. In 2003, Chien-Jan proposed a nonce-based authentication scheme using a smart card [7]. The scheme provides for the user and server to mutually authenticate one another. However, it is necessary to set up a verification table in this scheme. A legitimate user cannot update his password conveniently and freely when its security faces potential threats. In 2004, Juang proposed an authentication scheme that provides a key agreement function [8]. In Juang's authenticated key agreement scheme using a smart card, the smart card has to compute the modular exponential operations to attain the shared session key. This may overload the smart card capability.

In 2009, Liao et al. proposed an improved scheme [9], enhancing the efficiency and the functionality of Hwang-Lee-Tang, Chien-Jan and Juang's scheme. Liao et al.'s scheme adopted a transformed identity (TID) to avoid identity duplication. Howerever, despite a claimed proof of security, Liao et al.'s scheme is insecure in the presence of an active adversary. This paper presents that an adversary may exploit TID to achieve offline guessing attack. The scheme is also exposed to man-in-the-middle attack. Their claimed theorems and proofs are incorrect. We detail the analysis of flaws in the scheme and its security proof. This paper proposes an improved scheme to resolve these weaknesses and preserve user anonymity, a crucial issue in e-commerce applications.

The remainder of this paper is organized as follows. In Section 2, we review Liao et al.'s scheme and present our attacks on Liao et al.'s scheme; thus invalidating the claimed security of the scheme. In Section 3, we demonstrate our proposed scheme. In Section 4, we analyze the security of our scheme. Finally, we conclude this work in Section 5.

2 Review of Liao et al.'s Scheme

This section reviews Liao et al.'s scheme. Notation is provided. Then, the registration, login and authentication, key agreement, password update phase of their scheme are described in turn.

- ID_u : the identity of user U.
- PW_u : the password of U.
- TS_u : U's registration time.
- h : a secure hash function.
- TID_u : U's transformed identity,
- x : the permanent secret key of server S.
- p : a large prime positive integer.
- g : a primitive element in Galois field $GF(p)$.
- n_u, n_s : random numbers generated by U and S, respectively.

- E_k, D_k : symmetric encryption/decryption functions using symmetric key k satisfying $D_k(E_k(m)) = m$.
- SK_u, SK_s : session keys generated by U and S, respectively. If the scheme ends successfully, then $SK_u = SK_s$.

Registration Phase. This phase is invoked once, when U initially registers to S, and is described as follows:

1. U submits the registration request $\langle ID_u, PW_u\rangle$ to S via a secure communication channel.
2. Upon receiving the registration request, S acquires the registration time TS_u and archives ID_u and related TS_u. Then, S computes the transformed identity $TID_u = TS_u||ID_u$, $A_u = h(TID_u \oplus x)$ and $B_u = (g^{A_u} mod\ p) \oplus PW_u$. Next, S stores TS_u, B_u and h in a smart card and issues the smart card to U.

Login and Authentication Phase. When U intends to login S, U and S need to mutually authenticate each other.

1. U connects his smart card to a reader. The smart card challenges U for U's ID_u and PW_u, then generates and stores a nonce n_u. Next, retrieve TS_u to generate the transformed identity, $TID_u = TS_u||ID_u$, compute $NTID_u = TID_u \oplus n_u$ and $C_u = h(B_u \oplus PW_u) \oplus n_u$. Finally, U sends the message $M_1 = \{ID_u, NTID_u, C_u\}$ to S.
2. After receiving the message M_1, S retrieves TS_u, corresponding to ID_u. If no such corresponding U matches, Sterminates the connection. Otherwise, S computes $TID_u = TS_u||ID_u$, $n'_u = NTID_u \oplus TID_u$, $A_u = h(TID_u \oplus x)$ and $g^{A_u} mod\ p$, then $h(g^{A_u} mod\ p)$ and $n''_u = C_u \oplus h(g^{A_u} mod\ p)$. If $n'_u = n''_u$, the received $NTID_u$ is truly sent from U and $n'_u = n''_u = n_u$. Hence, U is authenticated. S stores n_u. Otherwise, S terminates the connection. S creates a nonce n_s, computes $D_u = C_u \oplus n_u \oplus n_s$ and $NTID_s = TID_u \oplus n_s$. Then S sends the message $M_2 = \{D_u, NTID_s\}$ to U.
3. After receiving the message M_2, U computes $n'_s = NTID_s \oplus TID_u$ and $n''_s = C_u \oplus n_u \oplus D_u$. If $n'_s = n''_s = n_s$, S is authenticated. Otherwise, U terminates the communication. U keeps n_s and computes $M_3 = (C_u \oplus n_u)||(n_s + 1)$. Then, U sends the message M_3 to S. The parameter $n_s + 1$ is the response to S.
4. Since $B_u \oplus PW_u = g^{A_u} mod\ p$, $C_u = h(B_u \oplus PW_u) \oplus n_u = h(g^{A_u} mod\ p) \oplus n_u$. Thus, $C_u \oplus n_u = h(g^{A_u} mod\ p)$. So, $M_3 = (C_u \oplus n_u)||(n_s + 1) = h(g^{A_u} mod\ p)||(n_s + 1)$. S can easily extract $n_s + 1$ from M_3 and find n_s in there. At this time, S ensures that U has the nonce, n_s.

The key agreement phase. After receiving n_s sent from S, U creates a session key $SK_u = h((B_u \oplus PW_u)||n_s||n_u)$. Once S ensures that U has n_s, it generates a session key $SK_s = h((g^{A_u} mod\ p)||n_s||n_u)$. Since $B_u = (g^{A_u} mod\ p) \oplus PW_u$, key agreement is achieved and the session key for the session communication is

$$SK_u = SK_s = h((B_u \oplus PW_u)||n_u||n_s) = h((g^{A_u} mod\ p)||n_s||n_u).$$

The password update phase. When U wants to change password, U inserts the smart card into a reader, announces a password update request at U's terminal and keys PW_u. Then the smart card calculates $B_u \oplus PW_u$ and U gives a new password $PW_u{}^*$. Finally, the smart card calculates $B_u{}^* = (B_u \oplus PW_u) \oplus PW_u{}^*$ and replaces B_u with this new $B_u{}^*$.

2.1 Weaknesses of Liao et al.'s Scheme

We point out security weaknesses of Liao et al.'s scheme. A smart card is a memory card with an embedded micro-processor to perform required operations specified in a scheme. No existing smart cards can prevent the information stored in them from being extracted, for example, by monitoring their power consumption [10,11]. Some other reverse engineering techniques are also available to extract information from smart cards. Hence we assume that once a smart card is stolen by an adversary, all the information stored in it are known to the adversary.

Offline Guessing Attack. Suppose U's smart card is compromised by an adversary A. Then A knows all the information $\langle TS_u, B_u, h\rangle$ stored in the smart card. If A possesses communication messages between U and S, A can perform the following offline guessing attack directly without interacting with S.

1. Using eavesdropped and stored session message $M_1 = \{ID_u, NTID_u, C_u\}$, A can calculate $TID_u = TS_u||ID_u$ and obtains n_u by computing $n_u = NTID_u \oplus TID_u$, and computes $K = C_u \oplus n_u = h(B_u \oplus PW_u)$.
2. Then A can perform an offline password guessing attack to obtain PW_u by guessing a candidate password $PW_u{}'$ and computing $K' = h(B_u \oplus PW'_u)$. If $K' = K$, which implies $PW'_u = PW_u$, A has successfully guessed U's password. Otherwise, A tries another candidate password.

In Liao et al.'s scheme, authors introduced and adopted the transformed identity $TID_u = TS_u||ID_u$ to avoid identity duplication. However, A may exploit TID_u to achieve the offline guessing attack. Therefore, unlike the authors' claim, without knowing x or A_u, A can impersonate the legal user U freely using the above attack.

Man-in-the-middle attack. We assume that attacker A interposes the communication between U and S. The attack scenario is outlined in Fig. 1, where a dashed line indicates that the corresponding message is intercepted by A en route to its destination. A more detailed description of the attack is as follows:

1. In the login and authentication phase, when U sends the message $M_1 = \{ID_u, NTID_u, C_u\}$ to S, A intercepts the message M_1.
2. Using intercepted message M_1 and creating a nonce n_A, A computes $NTID_A = NTID_u \oplus n_A$ and $C_A = C_u \oplus n_A$, A forges a message :

$$M_1^* = \{ID_u, NTID_A, C_A\}.$$

   Then A sends the forged message M_1^*, as if it originated from U.

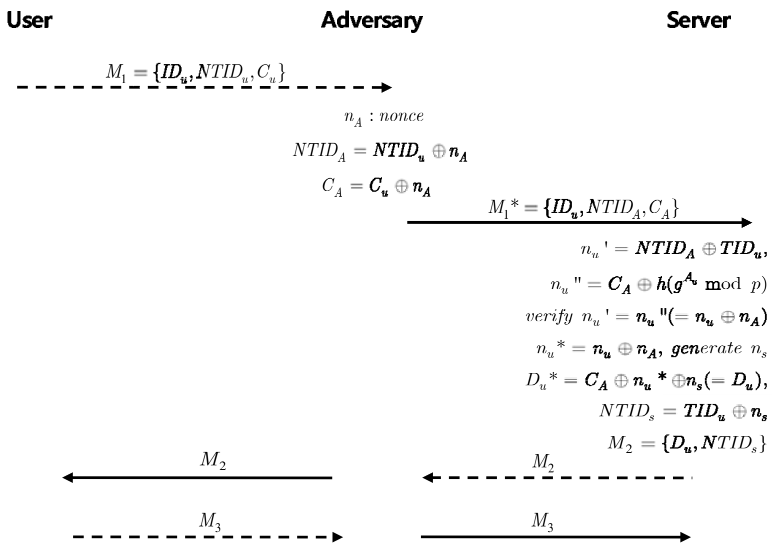

User
Adversary
Server
$M_1 = \{ID_u, NTID_u, C_u\}$
n_A : nonce
$NTID_A = NTID_u \oplus n_A$
$C_A = C_u \oplus n_A$
$M_1{}^* = \{ID_u, NTID_A, C_A\}$
$n_u{}' = NTID_A \oplus TID_u$,
$n_u{}'' = C_A \oplus h(g^{A_u} \bmod p)$
verify $n_u{}' = n_u{}''(= n_u \oplus n_A)$
$n_u{}^* = n_u \oplus n_A$, generate n_s
$D_u{}^* = C_A \oplus n_u{}^* \oplus n_s (= D_u)$,
$NTID_s = TID_u \oplus n_s$
$M_2 = \{D_u, NTID_s\}$
M_2
M_2
M_3
M_3

Fig. 1. Man-in-the-middle attack on Liao et al.'s scheme

3. According to Liao et al.'s scheme, upon receiving the message M_1^*, S retrieves TS_u, corresponding to ID_u, and computes $TID_u = TS_u || ID_u$, $n'_u = NTID_A \oplus TID_u$, $A_u = h(TID_u \oplus x)$, $g^{A_u} mod\ p$, then $h(g^{A_u} mod\ p)$ and computes $n''_u = C_A \oplus h(g^{A_u} mod\ p)$. Since $n'_u = n''_u$, the communication continues. In this situation, both n'_u and n''_u are equal to $n_u \oplus n_A$ where n_A is generated by A. Thus the forged message passes the verification test of S, S thinks U is authenticated. Furthermore, S keeps $n_u^* = n_u \oplus n_A$ at the server and generates a nonce n_s and computes $D_u^* = C_A \oplus n_u^* \oplus n_s$, $NTID_s = TID_u \oplus n_s$. Actually, $D_u^* = D_u = C_u \oplus n_u \oplus n_s$. Then S sends the message $M_2 = \{D_u, NTID_s\}$ to U. However, this message is intercepted by A.
4. A forwards the message M_2 to U. Then U operates, as specified in Liao et al.'s scheme, and sends the message $M_3 = (C_u \oplus n_u)||(n_s + 1)$ to S. However, this message is intercepted by A and forwards this message to S, as if it originated from U.
5. According to Liao et al.'s scheme, upon receiving the message M_2, S computes $B_u \oplus PW_u = g^{A_u} mod\, p$, $C_A = h(B_u \oplus PW_u) \oplus n_u^* = h(g^{A_u} mod\, p) \oplus n_u^*$. Thus $C_A \oplus n_u^* = C_u \oplus n_u = h(g^{A_u} mod\ p)$. Therefore, M_3 and computed $h(g^{A_u} mod\ p)||(n_s + 1)$ are equal. Since M_3 is valid, this passes, verifying U has n_s.

Following this, as described in the above attack, S will compute the wrong session key $SK_s = h((g^{A_u} mod\ p)||n_s||(n_u \oplus n_A))$. However, S cannot detect the generation of this wrong session key, because S authenticates U by verifying $n_u^* (= n_u \oplus n_A)$. From now, U and S shall use mutually different session keys in encrypting/decrypting their messages. Unlike Liao et al.'s security analysis, the forged messages made by A pass the verification test of U and S, because communicating parties check the validity of the received messages using the nonce. Liao et al.'s scheme cannot detect this attack and prevent communicating parties from maintaining the invalid sessions. Through this attack, A can make two parties believe and use an unintended session key.

Other weaknesses

1. In order to prove the correctness of mutual authentication, theorems 1 and 2 are provided with their proofs in the login and authentication phase of Liao et al.'s scheme [9]. In theorem 1, Liao et al. claim that if $n'_u = n''_u$, then U is authenticated. However, as previously described in our man-in-the-middle attack, A might intercept M_1 and forge a message M_1^* and forward M_1^* to S. Thus, unlike Liao et al.'s proof of theorem 1, proof of $n'_u = n''_u$ does not verify $NTID_u$ is really transmitted by U. Similarly, the proof of theorem 2 also does not mean S is authenticated. Therefore, from theorems 1 and 2, the correctness of the mutual authentication between U and S is not proven. As observed in our man-in-the-middle attack, mutual authentication is not achieved in Liao et al.'s scheme.
2. In Liao et al.'s scheme, when adversary A obtains transmitted messages between U and S, A can know who communicates with S. Recently, the authentication schemes are not only concerned about providing mutual authentication with key exchange, but also preserving user anonymity, because user privacy is an important issue in many e-commerce applications. Liao et al.'s scheme is also vulnerable to insider attack. It is obvious that S can launch an insider attack, which is undesirable, since in the registration phase, U sends the value PW_u to S.

3 Proposed Scheme

In this section, we propose an improved authentication scheme that resolves the security weaknesses described in the previous section. Fig.2 illustrates the scheme.

Registration Phase. The registration phase is invoked once, when U initially registers to S, and is described as follows:

1. U chooses ID_u and PW_u, generates a random number b, then computes $\alpha = h(b \oplus PW_u)$ and submits the registration request $\langle ID_u, \alpha \rangle$ to S via a secure communication channel.
2. Upon receiving the registration request, S acquires the registration time TS_u and archives U's ID_u and related $\beta = h(TS_u)$ for later use. Then S computes $TID_u = \beta || ID_u$, $A_u = h(TID_u \oplus x)$ and $B_u = (g^{A_u} mod\ p) \oplus \alpha$. Finally, stores the values β, B_u and h in a smart card and issues the smart card to U.
3. U enters b into the smart card, then U's smart card contains β, B_u, h and b. From now on U does not need to remember b.

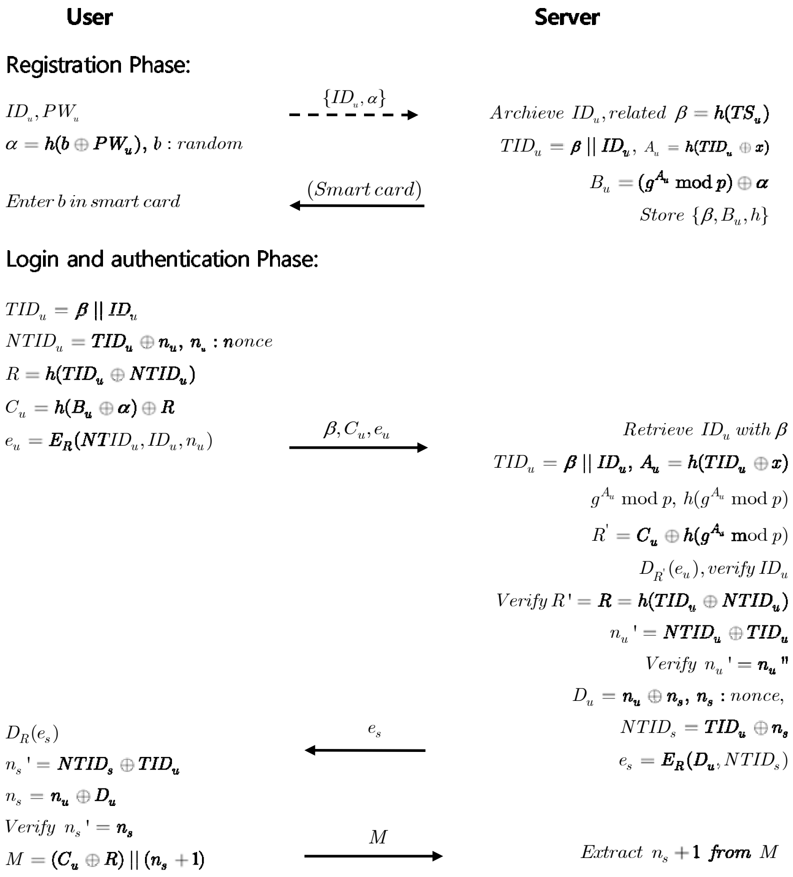

User
Server
Registration Phase:
ID_u, PW_u
$\alpha = h(b \oplus PW_u)$, b : random
$\{ID_u, \alpha\}$
Archieve ID_u, related $\beta = h(TS_u)$
$TID_u = \beta || ID_u$, $A_u = h(TID_u \oplus x)$
$B_u = (g^{A_u} \bmod p) \oplus \alpha$
Store $\{\beta, B_u, h\}$
(Smart card)
Enter b in smart card
Login and authentication Phase:
$TID_u = \beta || ID_u$
$NTID_u = TID_u \oplus n_u$, n_u : nonce
$R = h(TID_u \oplus NTID_u)$
$C_u = h(B_u \oplus \alpha) \oplus R$
$e_u = E_R(NTID_u, ID_u, n_u)$
β, C_u, e_u
Retrieve ID_u with β
$TID_u = \beta || ID_u$, $A_u = h(TID_u \oplus x)$
$g^{A_u} \bmod p$, $h(g^{A_u} \bmod p)$
$R' = C_u \oplus h(g^{A_u} \bmod p)$
$D_{R'}(e_u)$, verify ID_u
Verify $R' = R = h(TID_u \oplus NTID_u)$
$n_u' = NTID_u \oplus TID_u$
Verify $n_u' = n_u$"
$D_u = n_u \oplus n_s$, n_s : nonce,
$NTID_s = TID_u \oplus n_s$
$e_s = E_R(D_u, NTID_s)$
e_s
$D_R(e_s)$
$n_s' = NTID_s \oplus TID_u$
$n_s = n_u \oplus D_u$
Verify $n_s' = n_s$
$M = (C_u \oplus R) || (n_s + 1)$
M
Extract $n_s + 1$ from M

Fig. 2. Proposed scheme

Login and Authentication Phase. This phase is invoked whenever U intends to login S.

1. U connects his smart card to a reader. The smart card challenges U for ID_u and PW_u, which are selected at U's application. Then the smart card generates a nonce n_u and retrieves the stored β to generate the transformed identity $TID_u = \beta||ID_u$. Next, U's smart card computes $NTID_u = TID_u \oplus n_u$, $R = h(TID_u \oplus NTID_u)$ and $C_u = h(B_u \oplus \alpha) \oplus R$. Then, U encrypts $NTID_u, ID_u$ and n_u using R, yielding $e_u = E_R(NTID_u, ID_u, n_u)$. Finally, U sends (β, C_u, e_u) to S.
2. Upon receiving U's login request, S retrieves ID_u corresponding to β. If no such corresponding ID_u matches, S disconnects the connection. Otherwise, S computes $TID_u = \beta||ID_u$, $A_u = h(TID_u \oplus x)$, $g^{A_u} mod\ p$, $h(g^{A_u} mod\ p)$ and $R = C_u \oplus h(g^{A_u} mod\ p)$. Then, S gets $NTID_u, ID_u$ and n_u by decrypting $D_R(e_u)$, and verifies ID_u and $R' = R = h(TID_u \oplus NTID_u)$. Next, S computes $n_u' = NTID_u \oplus TID_u$. If $n_u' = n_u$, the received $NTID_u$ is

truly sent from U. Hence, U is authenticated. S stores n_u. Otherwise, S disconnects the connection. S creates a nonce n_s randomly and computes $D_u = n_u \oplus n_s$ and $NTID_s = TID_u \oplus n_s$. Then S, encrypts D_u and $NTID_s$ using R, yielding $e_s = E_R(D_u, NTID_s)$. Finally, U sends e_s to U.

3. After receiving e_s, U gets D_u and $NTID_s$ by decrypting $D_R(e_s)$. U computes $n_s' = NTID_s \oplus TID_u$ and $n_s = n_u \oplus D_u$. If $n_s' = n_s$, S is authenticated. U keeps n_s at U's terminal. Otherwise, U disconnects the connection. Next, U computes $M = (C_u \oplus R)||(n_s + 1)$. Then U sends M to S. The parameter $n_s + 1$ is the response to S.
4. Upon receiving M, since $C_u \oplus R = h(g^{A_u} mod\ p)$, $(C_u \oplus R)||(n_s + 1) = h(g^{A_u} mod\ p)||(n_s + 1)$. S can easily extract $n_s + 1$ from M and find n_s in there. At this time, S ensures that U has the nonce n_s.

The key agreement phase. After receiving the nonce n_s sent from S, U creates a session key $SK_u = h((B_u \oplus \alpha)||n_u||n_s)$. Once S ensures that U has the nonce n_s, it generates a session key $SK_s = h((g^{A_u} mod\ p)||n_s||n_u)$. Since $B_u = (g^{A_u} mod\ p) \oplus \alpha$ is computed in the registration phase, the key agreement is achieved and the session key for the session communication is

$$SK_u = SK_s = h((B_u \oplus \alpha)||n_u||n_s) = h((g^{A_u} modp)||n_s||n_u).$$

The password update phase. When U intends to change password, U inserts his smart card into a reader, announces a password update request at U's terminal and keys PW_u. Then, the smart card calculates $B_u \oplus h(b \oplus PW_u)$ and U gives a new password $PW_u{}^*$. Finally, the smart card calculates $B_u{}^* = (B_u \oplus h(b \oplus PW_u)) \oplus h(b \oplus PW_u{}^*)$ and replaces B_u with this new $B_u{}^*$.

4 Security Analysis

In this section, we briefly demonstrate that our proposed scheme is secure against an offline guessing attack, a man-in-the-middle attack, a stolen smart card attack and an insider attack.

1. Resistance to offline guessing attack. Suppose adversary A knows all the values (β, B_u, h, b) in U's smart card and intercepts $(\beta, C_u, e_u, e_s, M)$ transmitted between U and S. Even if A uses all the intercepted messages and extracted values in U's smart card, the offline guessing attack is impossible, because A cannot get R without knowing $g^{A_u} mod\ p$. Therefore, the proposed scheme is secure against offline guessing attack described in Section 2.
2. Resistance to man-in-the-middle attack. An adversary A may intercept or eavesdrop on the communication between U and S. After intercepting the message (β, C_u, e_u) sent by U, A may impersonate and replay the message to S. Even if A has the response message e_s from S, A cannot extract any values in e_s without knowing R which is never exposed on the communication. In addition, A cannot forge a message to impersonate U or S without knowing

Table 1. Functionality comparison of related schemes

	Proposed Scheme	[9]	[8]
User Anonymity	Yes	No	No
Communication/computation cost	Very low	Very low	Very low
Mutual authentication	Yes	No	Yes
Session key agreement	Yes	Yes	Yes

R. Using the symmetric key R, our proposed scheme prevents the man-in-the-middle attack described in Section 2. Moreover, R does not need to be exchanged during communication; U and S can get R by it computing on each side. Thus, the proposed scheme can withstand the man-in-the-middle attack.

3. Resistance to stolen smart card attack. Suppose A has stolen U's smart card and recorded the transmitted messages $(\beta, C_u, e_u, e_s, M)$ during one of U's past sessions. However, since A does not know ID_u and PW_u, A cannot forge message between U and S that passes login verification or forge SK_u and SK_s without knowing PW_u, n_u and n_s. Therefore, the proposed scheme can withstand the stolen smart card attack.
4. Resistance to insider attack. Since U registers to S by presenting $\alpha = h(b \oplus PW_u)$ instead of PW_u, the insider S cannot directly obtain PW_u. Furthermore, as b is not revealed to S, the insider of S cannot obtain PW_u by performing an offline guessing attack on α. Therefore, the proposed scheme can resist the insider attack.

In Table 1, we summarize the functionality comparison between our proposed scheme and the related schemes.

5 Conclusion

This work has considered the security of Liao et al.'s authentication scheme with key agreement. Although Liao et al.'s scheme claimed proof of its security, we show that the scheme is insecure against an offline guessing attack and man-in-the-middle attack and find flaws in the reasoning of the proof. We propose an improved scheme with better resistance to the offline guessing attack, man-in-the-middle attack, stolen smart card attack and insider attack to avoid these attacks.

References

1. Bellovin, S.M., Merritt, M.: Encryped key exchange: password-based protocols secure against dictionary attacks. In: IEEE Symposium on research in security and privacy, pp. 72–84. IEEE Computer Society, Los Alamitos (1992)
2. Botko, V., Mackenzie, P., Patel, S.: Provable secure password-authenticated key exchange using Diffie-Hellman. pp.156–171 (2000)

3. Jablon, D.P.: Strong password-only authenticated key exchange. ACM SIGCOMM Computer Communication Review 26(5), 5–26 (1996)
4. Wu, T.: The Secure Remote Password protocol. In: Internet Society Network and Distributed Systems Security Symposium (NDSS), pp. 97–111 (1998)
5. Yang, G., Wong, D.S., Wong, H., Deng, X.: Two-factor mutual authentication based on smart cards and passwords. Journal of computer and system sciences 74(7), 1160–1172 (2008)
6. Hwang, M.S., Lee, C.C., Tang, Y.L.: A simple remote user authentication scheme. Mathematical and Computer Modeling 36, 103–107 (2002)
7. Chien, H.Y., Jan, J.K.: Robust and simple authentication protocol. Computer Journal 46, 193–201 (2003)
8. Juang, W.S.: Efficient password authenticated key agreement using smart cards. Computers and Security 23(2), 167–173 (2004)
9. Liao, C.H., Chen, H.C., Wang, C.T.: An exquisite mutual authentication scheme with key agreement using smart card. An International Journal of Computing and Informatics (Informatica) 33(2), 125–132 (2009)
10. Kocher, P., Jaffe, J., June, B.: Differential power analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999)
11. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Examming smart card security under the threat of power analysis attacks. IEEE Transactions on Computer 51(5), 541–552 (2002)

Towards a Dynamic Federation Framework Based on SAML and Automated Trust Negotiation

Yicun Zuo, Xiling Luo, and Feng Zeng

School of Electronic and Information Engineering, Beihang University, China
yicunyi@gmail.com, luoxiling@buaa.edu.cn, zengf@adcc.com.cn

Abstract. One disadvantage with current Federated Identity Management systems is the establishment of the federation is based on a preestablished relying relationship between Service Provider and Identity Provider. The contribution of this paper is a proposal for the integration of Federated Identity Management with Automated Trust Negotiation to establish a Dynamic Federation, which makes the sharing of user information among potential business partners easier and more flexible, and provides better protection of users' privacy at the same time. In this paper, the architecture, main information exchange protocol and prototype implementation of Dynamic Federation Framework are described in detail.

Keywords: Federated Identity Management; SAML; Automated Trust Negotiation; Single Sign-On.

1 Introduction

Traditionally, authorization has been based on the identity of the entity requesting access to a resource, either directly or through roles [1]. However, when a requester and a server belong to different security domains controlled by different authorities, this approach is inconvenient.

To solve this problem, the concept of Federated Identity Management (FIdM) [2] has been brought forward. FIdM system allows the use of the same user's Personal Identification Information (PII) across multiple organizations within a federation [3]. The PII includes users' login names, user properties and user identity attributes. FIdM systems involve at least two types of entities: identity providers (IdP) and service providers (SP) [4]. An IdP manages user authentication and user-identity-relevant information. A SP offers services to users who satisfy the policy requirements associated with these services. It specifies and enforces the access-control policies for the resources it offers. Several existing FIdM systems are: SAML (Security Assertion Markup Language) [5,9], Liberty ID-FF (Identity Federation Framework) [6], and WS-Federation [7].

One disadvantage with current FIdM systems is that the trust between SPs and IdPs in one federation is preestablished at design time, which limits the cooperation with potential business partners. In other words, the SP's willingness to rely on information from an IdP depends on the existence of a trust relationship with the IdP[8],which may be based on a commercial agreement off line.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 254–262, 2010.

An approach to improve the above disadvantage is to integrate FIdM with automated trust negotiation (ATN) [1,4,10] to establish a dynamic federation, which makes the sharing of user information among potential business partners easier and more flexible, thereby promoting the commercial interest, and provides better protection of users' privacy at the same time.

In this paper, we proposed a Dynamic Federated framework (DFed) based on SAML and ATN, in which trust between SPs are established at runtime. And this is the main difference between DFed and other existing FIdM systems. In our framework, we don't differentiate between SPs and IdPs. An organization can act as both an IdP and a SP.

The rest of the paper is organized as follows. Section 2 indicates several important requirements for authorization systems designed for open systems. Section 3 and 4 present the details of the DFed framework architecture and Multi-Domain Single Sign-On protocol, respectively. In Section 5 we describe the implementation of a prototype system of DFed. Conclusions of this effort along with some future research plans are described in Section 6.

2 Design Requirements

Our goal of designing DFed is to provide a migration path for the integration of ATN technologies into existing FIdM systems, which meet the needs of open systems to the highest degree possible.

In large-scale open systems, business service providers often wish to realize the competitive advantages offered by allowing qualified outsiders access to their resources under certain conditions. One effective way to gain large number of users is to establish federation with business partners. Then the users' PII could be shared among SPs within the federation under users' permission. Given that there are many limitations of the traditional way to establish a federation mentioned in the above section, and there are compelling business reasons for resource providers to share user information with potential business partners, we can immediately recognize 5 important requirements:

- Bilateral trust establishment between different SPs. In open environment, we assume that there are no preexisting trust relationships between two SPs. For establishing a dynamic federation, it is important to allow these entities to establish trust relationships with one another at runtime.
- Federation of user attributes. Whenever the DFed are built up, SPs should have some means to make user's PII be shared.
- Support Multi-Domain Single Sign-On (SSO). User information sharing is the desire of E-Commerce Service Providers. For users, the ability of SSO provided by DFed will bring them more convenient when they surfing on the internet.
- Privacy preservation. To protect users' sensitive information, whenever possible, the users' sensitive information should not be shared to those that do not need them.
- Communication security. The communications between all interacting entities in DFed should be secured against external attackers.

3 The DFed Architecture

Figure 1 illustrates the DFed architecture. In this section we describe the main components that make up the architecture. This architecture is based upon SAML and ATN.

A DFed framework consists of many Dynamic Federation Service Providers (DFSP), which both perform the functionality of an IdP and an SP, containing the necessary components required to meet the requirements mentioned in part 2. DFSPs exchange information according to an on-demand dynamic protocol, as we detail later.

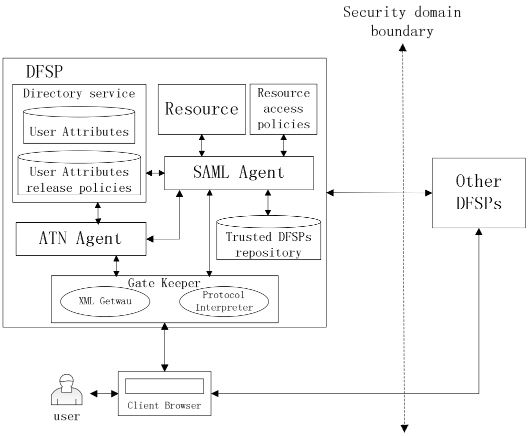


Fig. 1. The DFed architecture

3.1 Gate Keeper

Gate Keeper (GK) is a software component contains a protocol interpreter and a XML gateway. The protocol interpreter is responsible for carrying out the steps of the DFed SSO protocol (discussed in Section 4). Since most messages exchanged in our frame work are based on XML, the XML gateway identify the message type (ATN or SAML) and forward them to ATN Agent or SAML Agent after some necessary data format conversion.

3.2 Directory Service

Directory service is a centralized service to store user's attributes and attributes' release policies. These release policies state conditions under which the attribute protected by them can be disclosed to another entity such as other DFSPs. Conditions are constraints against the interacting entity's credentials and their attributes.

3.3 Trusted DFSPs Repository

The trusted DFSPs repository component is a database which stores the information of DFSPs which has established a trust relationship with the DFSP it belongs to.

There is a limit time, such as 15 days, for each record in the repository. If one DFSP doesn't have communication with local DFSP in 15 days, it will be deleted automatically. Of cause this limit time could be modified by the administrator in accordance with the requirements.

3.4 SAML Agent

The SAML Agent is responsible for understanding the SAML protocol. It includes three main components, SSO Service (SS), Identity Manager (IdM) Module and Service Manager (SM) Module, as shown in figure2.

The SSO Service is the first point of contact at the SAML Agent when the authentication request taking place. The SSO service tests whether the requesting party is a trusted DFSP according to the record in trusted DFSPs repository and initiates the authentication process at the IdM. It also has some means of interacting with ATN Agent.

The Identity Manager manages user credentials and attributes. Upon request the IdM will assert authentication statements or attribute statements to requesting parties, according to the result of the trust negotiation session with it.

The Service Manager manages secured resources. User access to resources is based on assertions received by the SM from an IdM (remote or local).

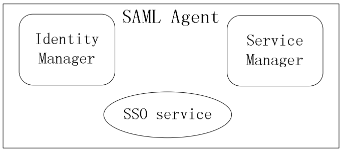


Fig. 2. SAML Agent architecture

3.5 ATN Agent

ATN Agent (AA) is responsible for understanding the ATN protocol and carrying out trust negotiation sessions on behalf of the DFSP that own it. In addition, an ATN Agent manages its owners' credentials and their corresponding release policies. The architecture of ATN Agent is shown in figure 3.

Credential Verifier verifies the content's integrity of every credential ATN Agent received by using a digital signature to guard against forgery. Furthermore, credential verifier must always check for expired and revoked credentials.

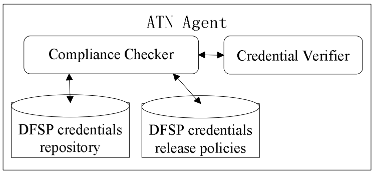

ATN Agent
Compliance Checker
Credential Verifier
DFSP credentials repository
DFSP credentials release policies

Fig. 3. ATN Agent architecture

Compliance Checker accepts a set of local DFSP credentials and a policy from remote counterpart. It returns a subset of these credentials that satisfies the policy. Or if there are one or more credentials protected by policies, it returns these policies.

4 Information Exchange Protocol

In this section, we present an overview of the information exchange protocol used in the DFed framework and discuss the way in which DFSP components interact during the execution of this protocol.

4.1 Communication Security

To fulfill requirement of communication security, Web Services (WS) messages with encryption and digital signatures can be used. Important communications between two DFSPs or between users and DFSPs, such as user logon session, SAML assertion delivery and trust negotiation message exchange, must occur inside a TLS or SSL tunnel used to provide confidentiality and integrity for the session.

4.2 SSO Protocol in DFed

As shown in figure 4, we present the SSO protocol of our DFed framework in a high level. This protocol takes place in 6 phases: (1) service request, (2) IdP select/URI input, (3) DFSP trust establishment, (4) user authentication, (5) user attributes negotiation and (6) response.

At the first phase, the user opens a web browser and intends to access the service located at DFSP2 with a HTTP request. The GK of DFSP2 (GK2) captures this request message and forwards it to the SM module (SM2) at DFSP2's SAML Agent. Since this user is not authenticated, SM2 answers with a web page which provides 3 optional operations: (1) logon or register directly; (2) select a DFSP from a list of DFSPs, which have established trust with DFSP2, as the IdP of the user; (3) choose other DFSPs as the IdP by inputting and submitting the URI of the DFSP, when the user is not a member of all DFSPs mentioned in operation (2). If the user chooses operation (1), the logon process will jump into phase 6. Here we don't elaborate on this situation.

We assume that the user is a member of DFSP1, and intend to choose it as the IdP during this logon process. DFSP1 is suggested to be an organization whose nature operations allows them to collect user's PII, such as banks and real-name social network sites, of course it is not limit to these organizations.

In phase 2, if DFSP1 is in the list, the user could select it directly, or the user should enter the URI of DFSP1. As soon as SM2 receive the information the user submitted, it sends an HTTP redirect response to the user browser. The Location HTTP header contains the destination URI of DFSP1 together with an <AuthnRequest> message which is a SAML request.

The GK of DFSP1 (GK1) identify the <AuthnRequest> is a SAML request and forwards it to the SSO Service (SS1) at DFSP1's SAML Agent. SS1 estimates whether there is priori trust between DFSP2 and DFSP1. That means they have built up a dynamic federation before. If it is, the logon process will jump into phase 4. Otherwise DFSP1 must establish trust in DFSP2 before providing SSO service to it.

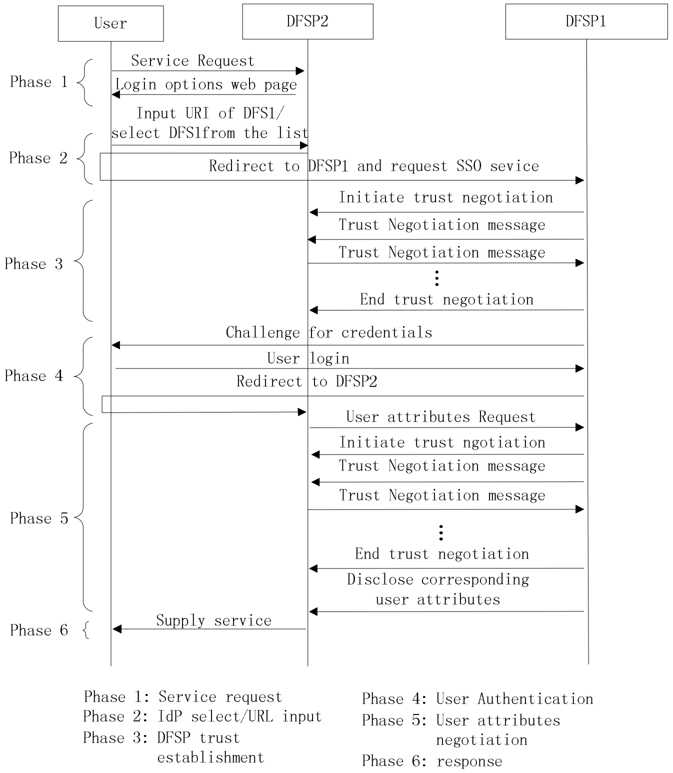

User
DFSP2
DFSP1
Phase 1
Service Request
Login options web page
Phase 2
Input URI of DFS1/ select DFS1from the list
Redirect to DFSP1 and request SSO sevice
Phase 3
Initiate trust negotiation
Trust Negotiation message
Trust Negotiation message
End trust negotiation
Phase 4
Challenge for credentials
User login
Redirect to DFSP2
Phase 5
User attributes Request
Initiate trust ngotiation
Trust Negotiation message
Trust Negotiation message
End trust negotiation
Disclose corresponding user attributes
Phase 6
Supply service
Phase 1: Service request
Phase 2: IdP select/URL input
Phase 3: DFSP trust establishment
Phase 4: User Authentication
Phase 5: User attributes negotiation
Phase 6: response

Fig. 4. The SSO Protocol of DFed

At the beginning of phase 3, DFSP1 initiates a trust negotiation session with DFSP2. The ATN Agent of DFSP1 (AA1) sends an Initiate Trust Negotiation (ITN) message to DFSP2. This ITN message serves as a flag to indicate that a new trust negotiation session is about to begin. After receiving an ITN message, GK2 will forward subsequent messages to ATN Agent of DFSP2 (AA2) for processing until an end trust negotiation message is received. AA1 then conducts an iterative exchange of trust negotiation messages with AA2. The DFSP credentials and corresponding policies are encoded in the body of these messages. Several rounds of Trust Negotiation messages may be required for both sides to determine whether an acceptable level of trust has been gained in the counterpart. At the end of this phase AA1 terminates this trust negotiation session by sending an end trust negotiation message to AA2. If the trust established successfully, DFSP2 and DFSP1 add information of the counterpart into their own trust DFSPs repository, respectively. Should the FSP1 and FSP2 fail to establish trust during this phase, DFSP2 reports a failure to the user after the connection between DFSP1 and DFSP2 is closed, and ask the user to choose other ways to logon.

When DFSP1 establish trust in DFSP2, the logon process enters phase 4. During this phase, SS1 determines whether the user has an existing logon security context at the IdM module (IdM1) at DFSP1's SAML Agent at fist. If not, IdM1 interacts with the user browser to challenge the user to provide valid credentials (such as <user name, password> pair, X.509 credentials, Etc.). The user provides valid credentials and a local logon security context is created for the user at IdM1. IdM1 checks the credentials against the directory of DFSP1. Then the user browser receive a redirect to SM2 from SS1 with a SAML <Response> message contains a SAML Assertion issued by IdM1 representing the user's logon security context.

SM2 can now determine whether this user shall have access to the resource it provide. It examines the resource access policy and asks IdM1 for some attributes of the user to satisfy this policy. At the beginning of phase5, SM2 sends IdM1 a SAML attribute request message which contains the SAML Assertion it received from the user just now. IdM1 first checks the assertion. If it matches with the assertion IdM1 generated recently, IdM1 knows which user it refers. Then IdM1 send a message to AA1 ask it initiate a new trust negotiation session with DFSP2 for user attribute exchange. This process is similar to phase3, but the guidance of this iterative exchange of trust negotiation messages is user attributes release policies. After the ending of this trust negotiation session, AA1 sends a message to inform IdM1 the result. If the value of <TN result> field in this message is TRUE, IdM1 answer SM2 with a new SAML assertion containing all attributes it requested. Otherwise, IdM1 answer SM2 with a refusal message.

In phase6, DFSP2creates a temporary user account use the attributes obtained from DFSP1 in phase5 and discloses the resource to the user. Or DFSP2 reports a failure to the user if the attribute request is refused.

5 Implementation

We have developed a prototype implementation of our Dynamic Federation Framework using the Java programming language.

The ATN Agent of DFSP is based on TrustBuilder[12], which currently supports the use of X.509[11] certificate for credentials and the IBM Trust Policy Language for policy specification, with future support for other policy languages and credential types. TrustBuilder has been successfully integrated with a number of protocols and applications, making it a good choice for use in our DFed Framework.

The SAML Agent of DFSP is based on Shibboleth [8], which is a open source software package for web SSO across or within organizational boundaries. It extends the SAML 2.0 SSO and attributes exchange mechanisms by specifying service-provider-first SSO profiles. The flexibility and scalability make it a suitable candidate for our DFed framework.

We choose OpenLDAP to provide directory service in our framework. OpenLDAP is an open source implementation of the Lightweight Directory Access Protocol [13]. It is a robust, commercial-grade and fully featured LDAP suite of applications and development tools.

We now comment on the performance of our implementation in a usage scenario. In this scenario, we developed two web sites act as DFSPs, which simulate online hotel reservation (hotel.example.com, Hotel for short) and airline tickets booking (airline.example.com, Airline for short), respectively. The user Alice is a member of

Airline and she'd like to reserve a room at Hotel. For accessing to the reservation system the following steps take place:

(1) Alice clicked the entrance of the reservation system at Hotel.

(2) Since Alice had not been logon, Hotel answered with a logon options web page.

(3) Since Alice was not a registered user of Hotel, and she was reluctant to comply the complex register stages, so she chose Airline to be the IdP during this logon process, and input the its URL (*airline.example.com*).

(4) Hotel redirected Alice to Airline with an authentication request.

(5) Since Hotel was a stranger to Airline, so Airline initiated a trust negotiation session with Hotel for the SSO service. Airline requested Hotel provide its full name, service type, metadata which would be used to configure a relying party at Airline, and prove it is a lawful organization.

(6) Above-mentioned information was willing to be released by Hotel without any condition except the metadata. It send a TN message to Airline include a operating credentials issued by official organization contains these information and a release policy of metadata.

(7) The metadata release policy stipulated that the release of metadata needs some information of the request party, including full name and official license. Airline was willing to release these information, so it sends its credentials contains these information to Hotel.

(8) Hotel requested Airline's metadata which will be used to configure asserting party at Hotel.

(9) Since Airline's metadata release policy has been satisfied by information released in (6), so Airline sent the metadata to Hotel.

(10) Airline ended this trust negotiation session.

(11) Airline challenged Alice to provide valid credentials.

(12) Alice log on.

(13) Airline redirected Alice to Hotel with a SAML assertion.

(14) Hotel connected to Airline and requested Alice's attributes to satisfy the resource access policy. These attributes include Alice's name, ID card number, age, sex, and cellphone number.

(15) Airline initiated a new trust negotiation session for user attributes exchange. Comply to Alice's attributes release policies, above-mentioned information will be only released to organizations who have a privacy protection credential issued by an accrediting organization.

(16) Hotel was willing to disclose this credential.

(17) Airline sent a SAML attribute assertion of Alice to Hotel.

(18) Hotel allowed Alice to access the reservation system.

6 Conclusions and Future Work

In this paper, DFed is described as a solution for internet service providers to establish federation with potential business partners at runtime. For providers, DFed framework could increase the user amount, and enhance the commercial interests. For users, it could bring more convenience with better protection of privacy. Future works will mainly be further research in the area of Legacy System integration and standardization. Since there are considerable heterogeneous systems on the internet, this work will be difficult but important.

Acknowledgments. The authors are grateful for the incisive and helpful comments and discussions of Chen Hailong, Huang Yan, Yan Ke, Sun Tianguang and many colleagues in the laboratory.

References

1. Winsborough, W.H., Li, N.: Towards practical automated trust negotiation. In: Proceedings of the Third International Workshop on Policies for Distributed Systems and Networks (Policy 2002), pp. 92–103. IEEE Computer Society Press, Los Alamitos (2002)
2. Shim, S., Bhalla, G., Pendyala, V.: Federated identity management. Computer 38(12), 120–122 (2005)
3. Suriadi, S., Foo, E., Josang, A.: A user-centric federated single sign-on system. Journal of Network and Computer Applications 32(2), 388–401 (2009)
4. Bhargav-spantzel, A., Squicciarini, A.C., Bertino, E.: Trust Negotiation in Identity Management. IEEE Security & Privacy 5(2), 55–63 (2007)
5. Hughes, J., Maler, E.: Security Assertion Markup Language (SAML) V2.0 Technical Overview. OASIS Working Draft 08 (2005)
6. Wason, T., Cantor, S., et al.: Liberty ID-FF Architecture Overview. Liberty Alliance (2004)
7. Bajaj, S., Della-Libera, G., et al.: Web Services Federation Language. WS-Federation (2003)
8. Cantor, S., et al.: Shibboleth Architecture: Protocols and Profiles. Internet2-MACE (2005)
9. Maler, E., et al.: Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS (2005)
10. Winsborough, W., Seamons, K., Jones, V.: Automated trust negotiation. In: DARPA Information Survivability Conference and Exposition, vol. 1, pp. 88–102 (2000)
11. Tuecke, S., Welch, V., Engert, D., Pearlman, L., Thompson, M.: Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile. RFC 3820 (2004)
12. Lee, A.J., Winslett, M., Perano, K.J.: TrustBuilder2: A Reconfigurable Framework for Trust Negotiation. In: IFIP Trust Management Conference (2009)
13. Hodges, J., Bob Morgan, R.L.: Lightweight Directory Access Protocol (v3). Technical Speci_cation (2002)

Research and Application of FlexRay High-Speed Bus on Transformer Substation Automation System

Hui Li, Hao Zhang, and Daogang Peng

College of Electric Power and Automation Engineering
Shanghai University of Electric Power
Shanghai, China

Abstract. Through researching FlexRay high-speed bus technology, this paper auxiliary builds electric power system hardware interface standards and software application layer standards. Meanwhile, this paper also realizes low cost and high speed communication network of digital transformer substation bottom layer, and cooperates with IEC61850 Standard to perform digital transformer substation network structure. The experiment result shows this paper's work has actual meaning to low end application such as intelligent building, family electric and industrial and mining enterprises.

Keywords: FlexRay, High-speed Bus, Transformer Substation Automation, Bottom Layer Communication.

1 Introduction

With the rapid development of electric power industry, the scale of electric power system enlarges continuously. System operation mode is more complex than before. That promotes development of automation technology of electric power system continually. Because of being difficult to realize complete test of all around characters of electromechanical protection devices, traditional electrical relay protection testing device is difficult to satisfy system requirement and adapt needs of technology development. For avoiding overlapping investment and implementing information share, we should specify electric power system and perform comprehensive consideration. Transformer substation automation implements secondary circuit simplification, data consistency and resource share, also improves management level and safety operation level together.

Transformer substation automation system is an automated system composed of multi computers and ICs, which replaces conventional measurement and monitor instruments, control panel, central signal system and remote panel. Using computer protection to replace conventional relay protection panel, it avoids disadvantage of conventional relay protection device lacking of communication with external devices. Therefore, transformer substation automation system is an application in transformer substation field combined with automation technology, computer technology and communication technology. Transformer substation automation system can acquire

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 263–270, 2010.

intact data and information, utilizing high performance of computing ability and logical judgment ability to monitor and control running and operation of various equipments in transformer substation. The system main characters include:

(1) Function Integration: According to the operation demand of transformer substation automation system, functions of secondary system will be comprehensively considered. Guiding with whole system design scheme, automation system performed optimized combination design to attain uniform coordination of relay protection and monitor system.
(2) Digitalized and modularized system structure: Digitalized protection, control and measurement device (Realization by computer and having digital communication ability) is helpful to connect various function modules through communication network and be convenient for interface function module expansion and information sharing.
(3) Operation and monitor on screen: When someone attended in transformer substation, connections between human and machine perform on back-stage machine. When no one attended in transformer substation, connections perform on host computer or workstation in remote scheduling center or operation control center.
(4) Intelligent running management: Mainly on unattended operation, human machine conversation and operation on screen, tabulation, printing, off-limit monitor and system information management, building real-time database and history data-base, OFF-ON operation and anti-misoperation lock, these character can reduce working staff's labor and perform some work which human unable to do.

Layered and distributed control system of Transformer substation automation system is shown in Figure 1. There are 3 layers in layered and distributed control system logically: station layer, bay layer and process layer.

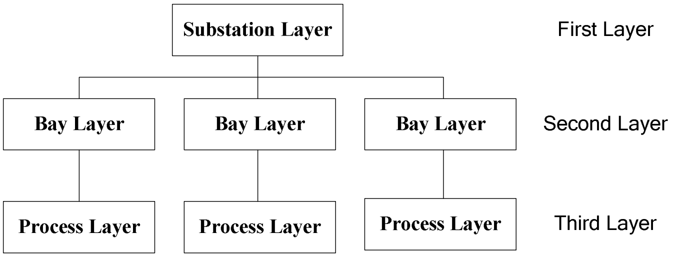


Fig. 1. Layered and Distributed Structure of Transformer Substation Automation System

From the top of structure, the first layer is substation layer. It acquires real-time data from bay layer and undertakes some functions in main control room, such as HMI, monitor, management and control between operator and remote monitor/maintain engineer station. It is also responsible to communicate with remote scheduling center.

The second layer is bay layer. It is responsible for communication management and control task of field devices and intelligent electronic devices of process layer.

Meanwhile, it is responsible for interpreting and exchanging work of communication standards.

The third layer is process layer. It acquires analog data, digital data and pulse data to protect and control operation output.

Therefore, Transformer substation automation system, through internal data communication, realizes information exchanging and sharing to reduce duplication configuration of transformer substation secondary equipments and simplify the interconnection of various subsystems. It not only reduces duplication of investment, but also improves overall system security and reliability.

As the special environment and substation automation system requirements, transformer substation automation system data network should meet the following requirements:

(1) Rapid real-time response capability;
(2) High reliability;
(3) Excellent electromagnetic compatibility.

2 Outline of FlexRay Technology

Association of FlexRay bus system early began in 1999 to the related content of cooperation. FlexRay Union officially was borned in 2000. In 2004, FlexRay issued the first specification and related tools. FlexRay Union's goal is to jointly establish a new type of high-quality communications network. It contains an excellent communications system, serial communication protocol details of the transceiver, consistent hardware and software interface specifications, and serves in the vehicle communication network development, production, and implementation process.

FlexRay bus system is a next-generation bus technology. Specifically, its communications system is characterized in the following areas:

(1) Bandwidth: FlexRay's bandwidth is not limited by protocol mechanisms. It can communicate with the fastest rate of 10Mbps. When using dual-channel redundant system, the rate will be up to 20Mbps, much higher than CAN bus.
(2) Scalability: FlexRay can use single and dual-channel modes, and realize mixed configuration.
(3) Flexibility: FlexRay network topology can use various modes, including point to point topology, active-passive bus topology and star topology. The physical layer device can use cable or fiber optic cable. Its communications data includes static segment and dynamic segment. FlexRay frame ID is corresponding to slot number, and also expressed the sender address. These are flexible communication mechanism performance of FlexRay.
(4) Certainty: FlexRay static segment strictly is based on time-triggered bus access method, while the dynamic segment can use the limited certainty of flexible time-triggered bus access. FlexRay bus is a trigger timing network system. Any network activities are arranged within the specified time slices. After the arrangement, they can not be changed. Therefore, FlexRay bus will never appear information overload.

(5) Clock synchronization: FlexRay network has an overall clock, and each control unit has a local clock. FlexRay system has a specific control algorithm, so that each individual node in the network realizes local clock synchronization with the overall clock, by means of offset correction and time correction method.

FlexRay and CAN bus performance comparison is shown in Table 1. FlexRay has a clear advantage in the baud rate, communication mode and arbitration compared to the CAN bus, but there are restrictions on the number of nodes. For example, when using the bus structure, it is only permitted for 22 nodes.

Table 1. FlexRay and CAN Performance Comparison

Performance Index	CAN	FlexRay
Baudrate	10Kbit/s-1Mbit/s	10Mbit/s
Communication	Event Drive	Time/Event Drive
Arbitration	CSMA/CD	TDMA
Frame ID	11/29 bits	11 bits
Transmission Medium	Cable / Fiber	Cable / Fiber

3 FlexRay Network Topology Structure

FlexRay has three kinds of network topology: bus, star and hybrid. The maximum number of these three types can support 22, 64 and 64 nodes. Each type can realize single channel and dual channel. In star topology, there is a method of conjunction class. FlexRay node to node distance and numbers of network nodes are limited in each type. The maximum distance between any two nodes is 24m and the maximum nodes are up to 22.

Figure 2 shows a passive bus connection network. IStubN means the distance from node to bus cross joint. ISpliceDistance means the distance between bus cross joints. In case of network nodes beyond two, extra nodes should be suspended.

Figure 3 shows the active star network topology. IActiveStarN means the distance between node N and active star connector. This topology introduces an active star linker. As a router, it transmits received information to other nodes. The advantage is when detecting the problem of illicit abnormal slip, it takes off the slip road initiatively to protect the other branch of normal communication. Furthermore, since

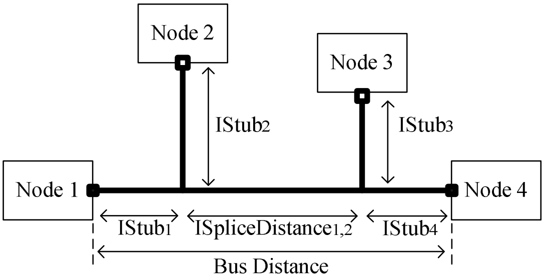


Fig. 2. FlexRay Bus Node Connection

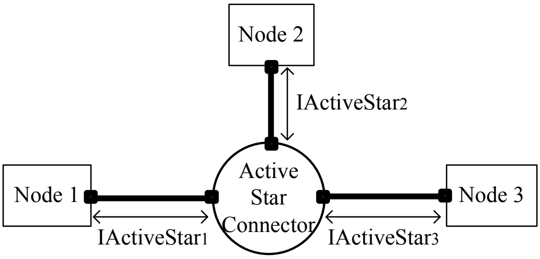


Fig. 3. FlexRay Active Star Node Connection

the active star connector can connect to other passive nodes, and they can also be connected together, which will extend system performance.

Hybrid FlexRay network topology usually combines with two or more topologies simultaneously. In FlexRay bus system, hybrid topology uses bus and star together, but the difficulty is to increase the network configuration.

4 FlexRay Applied to Communication Simulation System Structure of Transformer Substation

Based on FlexRay high-speed bus, the communication simulation system of transformer substation is shown in Figure 4:

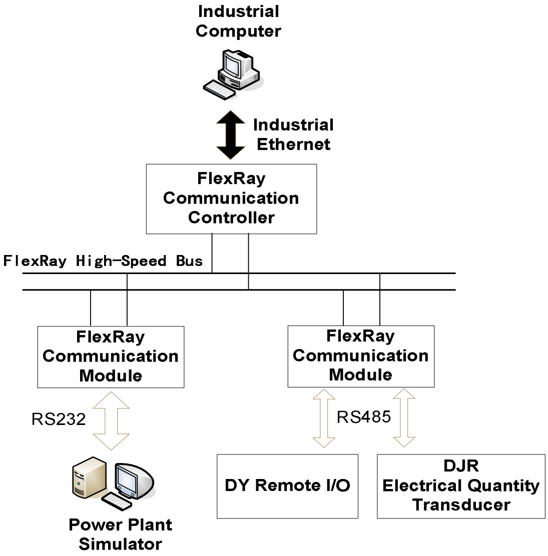
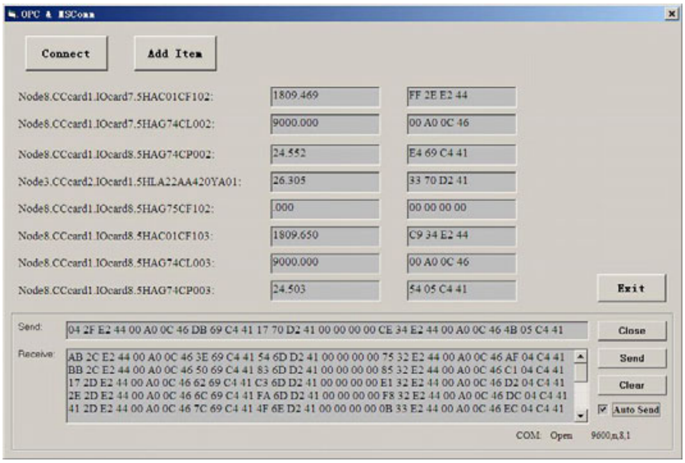


Fig. 4. Transformer Substation Communication Simulation System Based on FlexRay

In Figure 4, the simulation system data are originated from power plant simulator and field data acquisition devices (including DY remote I/O and DJR electrical quantity transducer). The electrical part of power plant simulator is used as simulation data source of transformer substation. On one hand, FlexRay communication module sends data through FlexRay high-speed bus to industrial computer, which acquires from power plant simulator. On the other hand, FlexRay communication module through RS485 bus acquires voltage, current, power, temperature, pressure and other data from DY remote I/O and electrical quantity transducer, and sends through FlexRay high-speed bus to monitor and control computer. FlexRay communication module and FlexRay communication controller are connected between two FlexRay high-speed buses. Industrial computer is responsible for remote monitor and control.

5 Experiment Result

The OPC client, which is on the side of power plant simulator, the screenshot of reading data software interface is shown in Figure 5. These data are sent through the RS232 interface to FlexRay communication modules:



Node8.CCcard1.IOcard7.5HAC01CF102:	1809.469	FF 2E E2 44
Node8.CCcard1.IOcard7.5HAG74CL002:	9000.000	00 A0 0C 46
Node8.CCcard1.IOcard8.5HAG74CP002:	24.552	E4 69 C4 41
Node3.CCcard2.IOcard1.5HLA22AA420YA01:	26.305	33 70 D2 41
Node8.CCcard1.IOcard8.5HAG75CF102:	.000	00 00 00 00
Node8.CCcard1.IOcard8.5HAC01CF103:	1809.650	C9 34 E2 44
Node8.CCcard1.IOcard8.5HAG74CL003:	9000.000	00 A0 0C 46
Node8.CCcard1.IOcard8.5HAG74CP003:	24.503	54 05 C4 41

Send: 04 2F E2 44 00 A0 0C 46 DB 69 C4 41 17 70 D2 41 00 00 00 00 CE 34 E2 44 00 A0 0C 46 4B 05 C4 41

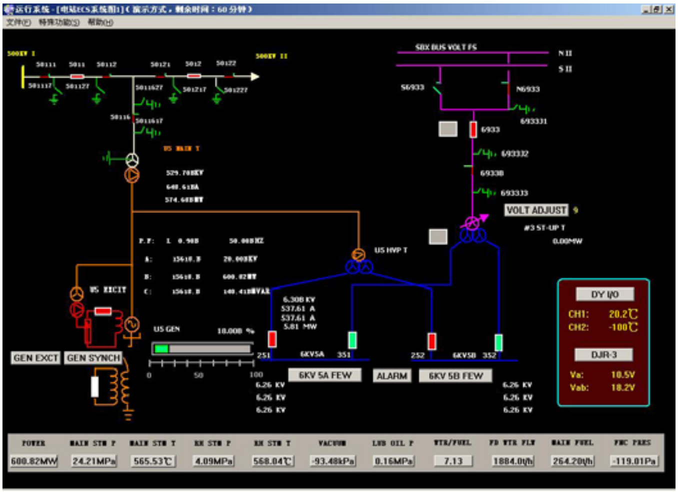
Receive:

```
AB 2C E2 44 00 A0 0C 46 3E 69 C4 41 54 6D D2 41 00 00 00 00 75 32 E2 44 00 A0 0C 46 AF 04 C4 41
BB 2C E2 44 00 A0 0C 46 50 69 C4 41 83 6D D2 41 00 00 00 00 85 32 E2 44 00 A0 0C 46 C1 04 C4 41
17 2D E2 44 00 A0 0C 46 62 69 C4 41 C3 6D D2 41 00 00 00 00 E1 32 E2 44 00 A0 0C 46 D2 04 C4 41
2E 2D E2 44 00 A0 0C 46 6C 69 C4 41 FA 6D D2 41 00 00 00 00 F8 32 E2 44 00 A0 0C 46 DC 04 C4 41
41 2D E2 44 00 A0 0C 46 7C 69 C4 41 4F 6E D2 41 00 00 00 00 0B 33 E2 44 00 A0 0C 46 EC 04 C4 41
```

COM: Open 9600,n,8,1

Fig. 5. OPC Client Reading Data from Power Plant Simulator

Monitor and control screenshot of industrial computer interface is shown in Figure 6. In configuration software Force-Control 6.0, we redraw the operation simulator interface of power plant simulator. The data in Figure 6 are power plant simulator actual operating data. DY remote I/O and DJR electrical quantity transducer power control interface are added on the right side of this figure. We can see that industrial computer via high-speed FlexRay bus can monitor various real-time data acquired from the field.



POWER 600.82MW
MAIN STM P 24.21MPa
MAIN STM T 565.53℃
RH STM P 4.09MPa
RH STM T 568.04℃
VACUUM -93.48kPa
LUB OIL P 0.16MPa
WTR/FUEL 7.13
FD WTR FLW 1884.0t/h
MAIN FUEL 264.20t/h
FWC PRES -119.01Pa

Fig. 6. Monitor and Control Interface of Industrial Computer

6 Conclusion

This paper researches on digital substation architecture in-depth, then designs digital transformer substation bottom layer communication network and puts forward specific implementation. As to the base of power plant simulator, this paper contructs the example of the bottom layer communication network simulation system based on FlexRay bus. FlexRay communication module on one hand acquires data from the power plant simulator and field data acquisition devices, on the other hand sends data to monitor and control computer. The host monitor and control computer realizes remote monitor and control through acquired data.

The rapid development of power industry has been expanding the scale of electricity system. System operation mode is more complex and power system automation requirement has gradually increased to promote the continuous development of power system automation technology. Conventional electrical relay protection test equipment can not meet the system requirements. It is difficult to achieve protection against mechanical and electrical characteristics of all aspects of comprehensive testing, and therefore it no longer needs to adapt to technological development. The paper simplifies the secondary circuit of transformer substation automation, also realizes data consistency, resource sharing and improves transformer substation operation and management level.

Acknowledgments. This work was sponsored by "Chen Guang" project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation(Grant No. 09CG61), Program of Shanghai Subject Chief Scientist(Grant No. 09XD1401900), and Science Foundation for The Excellent Youth Scholars of Shanghai Municipal Education Commission(Grant No. sdl09001).

References

1. Ri-cai, G., Zi-zhi, X., Xin-qian, X.: Research and Application of Typical Design for 220kV and 110kV Substations. Power System Technology 31(6), 23–30 (2007)
2. Seok, K., Kweon, Y., Shin, K.G.: Statistical Real-Time Communication over Ethernet. IEEE transactions on parallel and distributed systems 14(3), 322–335 (2003)
3. Huiyu, L.I., Shenglin, Y.U., Kan, J.I., Xiaoli, L.: A Layered and Distributed Dual-bus Communication Scheme for Substation Automation Systems Based on Point-to-point Communication. Automation of Electric Power Systems 31(16), 99–102 (2007)
4. Peng, D., Zhang, H., Yang, L., et al.: Research of Remote Condition Monitoring System for Turbo-generator Unit Based on B/S Model. In: Proceedings of 2008 International Conference on Condition Monitoring and Diagnosis, Beijing. China, April 21-24, pp. 175–179 (2008)
5. Kim, W.S., Kim, H.A., Ahn, J.-H., Moon, B.: System-Level Development and Verification of the FlexRay Communication Controller Model Based on System. In: Proceedings of 2008 Second International Conference on Future Generation Communication and Networking, pp. 124–127 (2008)

A Task Scheduling Algorithm Based on Load Balancing in Cloud Computing

Yiqiu Fang[1], Fei Wang[1], and Junwei Ge[2]

[1] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, 400065 Chongqing, China
[2] College of Software, Chongqing University of Posts and Telecommunications, 400065, Chongqing, China
fangyq@cqupt.edu.cn, wangfei___zxc@163.com, gejw@cqupt.edu.cn

Abstract. Efficient task scheduling mechanism can meet users' requirements, and improve the resource utilization, thereby enhancing the overall performance of the cloud computing environment. But the task scheduling in grid computing is often about the static task requirements, and the resources utilization rate is also low. According to the new features of cloud computing, such as flexibility, virtualization and etc, this paper discusses a two levels task scheduling mechanism based on load balancing in cloud computing. This task scheduling mechanism can not only meet user's requirements, but also get high resource utilization, which was proved by the simulation results in the CloudSim toolkit.

Keywords: cloud computing; task scheduling; virtualization; load balancing.

1 Introduction

As the key and frontier field of the current domestic and international computer technology, cloud computing is a computing paradigm that can provide dynamic and scalable virtual resources through the Internet service to users on demand, and also it is further development of distributed computing, parallel computing and grid computing [1]. Its main advantage is that it can quickly reduce the hardware costs and increase computational power and storage capacity; users can access high quality of service by low cost. And it is unnecessary to purchase expensive hardware equipment, as well as upgrade and maintain frequently.

Task scheduling is an important part of cloud computing, which is a mechanism that maps users' tasks to appropriate resources to execute, its efficiency will directly affect the performance of the whole cloud computing environment. Compared with grid computing, there are many unique properties and the mainly include virtualization and flexibility for cloud computing [2]. By using virtualization technology, all physical resources are virtualized and are transparent to user. All user has their own virtual machine and don't affect each other, which is created according to the user's requirement. Furthermore, one or multiple virtual machines can be run on a single host, and the utilization of resources is improved effectively, and the running independency of users' application is ensured as well as the information security of system and service availability is improved. Flexibility is the resource

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 271–277, 2010.

provided by cloud computing environment which can be increased or reduced dynamically according to users' demand for tasks. Owing to these new features, the task scheduling mechanism for grid computing can not work effectively in the cloud computing environment.

In this paper, a task scheduling mechanism based on the two levels of load balance is discussed, which consider the flexibility and virtualization in cloud computing to meet the dynamic task requirements of users and improve the utilization of resources.

2 Scheduling Model

Cloud Computing Architecture includes three layers, application layer, platform layer and infrastructure layer [3].The application layer is oriented to users, it implements the interaction mechanism between user and service provider with the support of platform layer. Users can submit tasks and receive the results through the application layer in the task scheduling process. The infrastructure layer is a set of virtual hardware resources and related management function, in which the implementation of internal process automation and optimization of resource management can provide optimal dynamic and flexible external infrastructure services. Furthermore, the platform layer is a set of software resources with versatility and reusability, which can provide an environment for cloud application to develop, run, manage and monitor. According to the above architecture, two levels scheduling model [4] are adopted in this paper as shown in Fig.1.

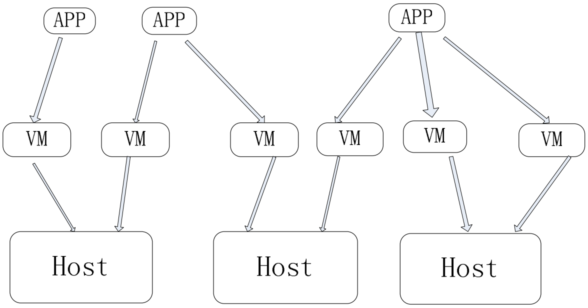


Fig. 1. Two levels scheduling Model in cloud computing

As shown in Figure 1, the first level scheduling is from the users' application to the virtual machine, and the second is from the virtual machine to host resources. In this two levels scheduling model, the first scheduler create the task description of a virtual machine, including the task of computing resources, network resources, storage resources, and other configuration information, according to resource demand of tasks. Then the second scheduler find appropriate resources for the virtual machine in the host resources under certain rules, based on each task description of virtual machine. Via the two levels scheduling, the task can obtain the required resources, and it would not lead to the resource allocated to some tasks is less than requirement and increase the implemental time while others are more than their requirements and lead to the waste of resources.

In order to describe the two levels scheduling model, task model and host resource model is established. All tasks in this paper are computational ones, only the Meta task is considered, and the tasks are independent each other, and the execution of the task replication is also not considered. The task model is described as follows:

The set of tasks is $T=\{t_0,t_2,\cdots,t_{n-1}\}$, and the number of tasks is $n=|T|$.The T , $t_i(i\in[0,n-1])$ indicates task i , $t_i=\{tId,tRr,tSta,tData,tVmch,tVm\}$, each property is defined as follows:

1) The task identification is defended as tId .

2) The required resource of one task is defined as $tRr=\{tC_0,tC_2,\cdots,tC_{k-1}\}$, where k is the number of resource type, $tC_j=\{j\in[0,k-1]\}$ is the ability of each resource.

3) The state of task is defined as $tSta$, where $tSta=\{tFree,tAllo,tSche,tWait,tExec,tComp\}$.

4) The relative data of task is defined as $tData$, including the computational amount tC , the input data tI and the output data tO , etc.

5) The description of virtual machine for a task is defined as $tVmch=\{tId,tRr,tData\}$, including the task identification tId , the required resource tRr and the related data $tData$.

6) The virtual machine of task is defined as $tVm=\{tId,thId\}$, where tId is the identification of virtual machine and $thId$ is the host identification allocated by the virtual machine.

The host holds all the physical resources for the task implementation, including computational resources, storage resources, network resources, and other hardware devices and the model is described as follows:

The set of host resources is defined as $H=\{h_0,h_1,\cdots,h_{m-1}\}$, and the size of set is $m=|H|$. In the H , $h_j(j\in[0,m-1])$ indicates Host j , $h_j=\{hId,hTcap,hFcap,hData\}$, each property is defined as follows:

1) The identification of host resource is defined as hId .

2) The total service ability that the host resource can provide is defined as $hTcap=\{hTc_0,hTc_1,\cdots,hTc_k\}$, where k is the number of resource type, $hTc_j(j\in[0,k-1])$ is the service ability of each resource.

3) The available used resource of host is defined as $hFcap=\{hFc_0,hFc_1,\cdots,hFc_k\}$.

4) The relative data of host is defined as $hData$, including the input bandwidth hIB and the output bandwidth hOB .

In this scheduling model, because the task can change dynamically, including the task arrival time and the demand for resources, which requires the corresponding virtual machine can change dynamically according to the change of task to deploy the resource in order to meet the dynamic demand. In this process, if the virtual machine requires additional resources and they can not be provided by the host. Then the virtual machine migration technology will be used. Two migration strategies can be taken: the first one is to migrate the virtual machine to a host and allocate the additional resources for the machine, and another one is to migrate the other virtual machines out of this host to vacate the free resources for the virtual machine.

3 Scheduling Algorithm

The load of virtual machine discussed in this paper is expressed by the predicted executing time of tasks running in the virtual machine, named as VL_i [5].And the load of host is expressed by the average load of the virtual machine that run on it, named as HL_i, it is defined as: $HL_i = \frac{\sum_{j=1}^{n} VL_j}{n}$, where the n is the number of virtual machines that run on the host.

From the HL_i, the average load value $avgl$ and load balancing evaluating value B of cloud computing environment can be defined as follows:

$$avgl = \frac{\sum_{i=1}^{m} HL_i}{m} \tag{1}$$

$$B = \frac{\sqrt{\sum_{i=1}^{m} (L_i - avgl)^2}}{m} \tag{2}$$

In the equations above, the number of hosts is m, the smaller value B the better load balancing and the bigger value B the worse load balancing.

In order to meet users' requirements and increase the utilization of resources, a scheduling algorithm based on load balancing is proposed in this paper. The algorithm is based on the former scheduling model discussed, considering the flexibility and virtualization features of cloud computing, it is divided into two levels scheduling, one is the mapping from task to a virtual machine, another is mapping from the virtual machine to host resources. Generally, for the requirement of the task, users want to get the best response time. Therefore, only task response time and the demand for resources are considered in this paper. At the same time, because tasks are dynamic, they may arrive randomly. If the tasks arrive at same time, they will be sorted ascending according to the resource applied by users. And if they arrive at different time, they will be sorted according to the time sequence arrived. The steps of this algorithm are described as follows:

Step1: According to the host resource model, establish the host resource set $H = \{h_0, h_2, \cdots, h_{m-1}\}$ and sort ascending order of their processing power.

Step2: According to the task model, establish the task set $T = \{t_0, t_2, \cdots, t_{n-1}\}$ .In this process, the first level scheduler establish the virtual machine description according to the properties of task ,providing configuration information for allocation of resources and creation of the virtual machine.

Step3: According to the virtual machine description of Task $t_i \in T$, select a host resource h_j that can meet the required resources and the load is lightest. If the host exists, create the virtual machine and allocate the required resource for it, then update the available resources $hFcap$ of Host h_j, otherwise take the Task t_i to the tail of the task queue and waiting for the next scheduling.

Step4: If the resource requirements of the Task t_i increase, find whether the host whose virtual machine of Task t_i run on can meet the additional required resources, if it exists, allocate the additional required resources for it, reconfigure the virtual machine, and then update the host's available resources. Otherwise, the virtual machine is migrated to the host with lightest load and the additional required resources to execute continuously.

Step5: If the resource requirements of the Task t_i reduce, release the excess resources that the virtual machine occupied, and update the available resources hold by the host.

Step6: If Task t_i has been completed, then destroy the virtual machine of Task t_i and release the occupied resources for the other unfinished tasks.

Step7: Calculate the load balancing evaluating value B in current environment, if B is greater than the threshold value B_0, that indicates the load balancing state is worse, select a virtual machine with lightest load and migrate it to the host which can meet the resource requirement with the lightest load.

Step8: Repeat step3 to 7 until all tasks are completed.

In the above algorithm, the virtual machine is scheduled to the host with lightest load each time. The advantage is to avoid overloading for the host hold more resources. If the current virtual machine is scheduled to a host, as the computational amount increase, leading to the virtual machine's load is heavy, resulting in load imbalance, then take the dynamic migration operation, keeping load balance in current environment.

4 Simulation Experiment and Result Analysis

In this paper, the CloudSim toolkit is used to simulate the scheduling algorithm discussed above [6]. In order to compared with the performance under different environments, a cloud computing environment and a grid computing environment with one hundred host node is set up to implement scheduling of one hundred to one thousand independent tasks, which demand can be dynamically changed during execution. This simulation mainly validates the advantage of the makespan and the resource utilization between this scheduling mechanism in cloud computing and the scheduling mechanism in paper [7] in grid computing, the expression of resource utilization are defined as follows:

$$RUsed = \left. \sum_{j=0}^{m-1} \frac{Rt_j}{Rc_j} \middle/ m \right. \tag{3}$$

In equation (3), the total execution time in host h_j is Rt_i, the actual occupied time in host h_j is Rc_j, and the number of host is m .The simulation result are described as Fig.2 and Fig.3:

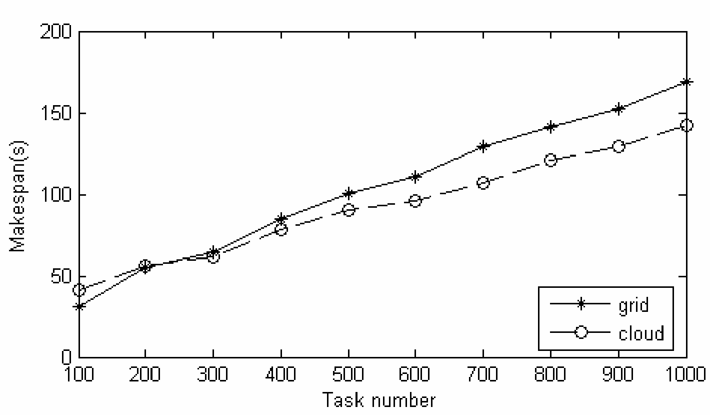


Fig. 2. Makespan Comparation

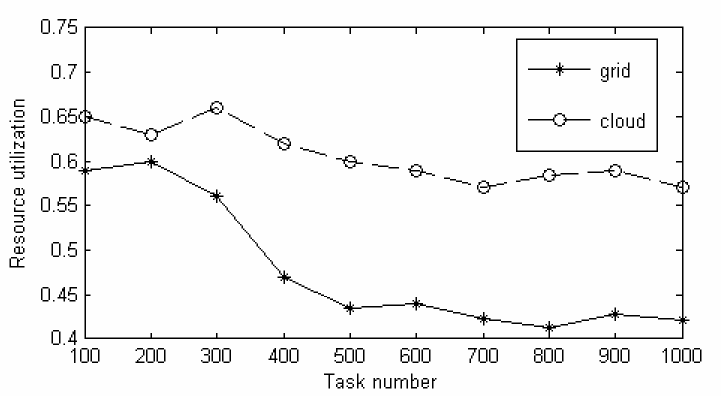


Fig. 3. Resource utilization Comparation

As shown in Figure 2, with the task number increasing, the makespan in cloud environment and grid environment are also increased. When the number of tasks and hosts difference is not large, the makespan in grid environment is lower than cloud environment. It is because the resources are allocated to a task is just it required in cloud computing, while the whole host resources are allocated to a task in grid computing. However, with the increase of ratio of number of tasks and hosts, the makespan in cloud environment is significantly less than it in grid environment, which is due to the flexibility in cloud computing, in which the resource can be allocate to the task dynamically stretching. As indicates in Figure 3, the resource utilization in cloud environment is significantly higher than grid environment and maintains a relatively stable level. The virtualization in cloud computing is considered fully in the scheduling mechanism for this paper, in which multiple virtual machines can be run in one host, when the load is imbalance, the dynamic migration approach is used to achieve a relatively load balancing and improve resource utilization.

5 Conclusions and Future Work

The traditional task scheduling in grid computing is to schedule the task directly to the host resources to execute, and it is not well to meet the dynamic requirements of users. A two-level scheduling mechanism based on load balancing is discussed in this paper. The scheduling mechanism take into account the dynamic requirements of users and the load balancing in cloud environment, therefore, it meets the dynamic requirements of users and increases the resource utilization. Through the CloudSim toolkit, we simulate this scheduling mechanism, proving the well scheduling effect. In the future work, we will consider more users' requirements, such as bandwidth, cost, etc., to establish a precise description of model for the users' requirements, to further improve the scheduling mechanism.

Acknowledgment

This work was supported by Chongqing Municipal Education Commission, NO.KJ090519.

References

1. Rimal, B.P., Choi, E., Lumb, I.: A Taxonomy and Survey of Cloud Computing Systems. In: Fifth International Joint Conference on INC,IMS and IDC, vol. 218, pp. 44–51 (2009)
2. Foster, I., Zhao, Y., Raicu, I., Lu, S.: Cloud Computing and Grid Computing 360-Degree Compared. In: Grid Computing Enviroments Workshop, pp. 1–10 (2008)
3. Armbrust, M.: Above the Clouds: A Berkeley View of Cloud Computing. In: EECS Department, University of California, Berkeley (2009)
4. Sadhasivam Dr., S., Jayarani, R.: Design and Implementation of an efficient Two-level Scheduler for Cloud Computing Environment. In: International Conference on Advances in Recent Technologies in Communication and Computing, vol. 148, pp. 884–886 (2009)
5. Xiang-hui, P., Er-hu, Z., Xue-yi, W., Guang-feng, L.: Load balancing algorithm of multi-cluster grid. Computer Enineering and Applications 45(35), 107–110 (2009)
6. The CLOUDS Lab.CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services [EB/OL], `http://www.gridbus.org/cloudsim/`
7. Bing-han, L., Jun, H., Xiang, H., Qi, L.: Grid Load Schedule Algorithm Based on QoS. Computer Engineering 35(24), 96–98 (2009)

The Accuracy Enhancements of Virtual Antenna for Location Based Services

Fu Tao, Huang Benxiong, and Mo Yijun

The Department of Electronics and Information Engineering,
Huazhong University of Science and Technology,
430074 Wuhan, China
{Fu.Tao,Huang.Benxiong,Mo.Yijun,9566911}@qq.com

Abstract. Measurement report (MR) base methods are a kind of cell identifier (CI) base methods whose parameters could be extracted from MRs and the cell configuration database (CCD). They are utilized for location based services (LBS), and take advantages in wireless network optimization. In this paper, we suggest the enhancements of virtual antenna method which is the improvement of enhanced CI-RXLEV method. We divide the location procedure into 2 steps, the first step is the virtual antenna method, and the second step utilizes the results calculated by the first step for filtering cells. So the cells which are far from the UE and nearby the serving cell could be filtered more precisely. We also improved the MPPE criterion cited by the virtual antenna. A new parameter is imported to avoiding MPPE score approaching zero. The experiment results show that our new enhancements perform better than the original virtual antenna method, and the position accuracy is close to the fingerprints.

Keywords: location based service (LBS); measurement report (MR); enhance CI-RXLEV (CI-RXLEV-E) algorithm; virtual antenna; position error size (PES).

1 Introduction

Measurement report (MR) based methods are a kind of location method whose parameters are extracted from MRs and the cell configuration database (CCD). Though satellite based technologies, such as GPS and Compass, have high accuracies, they also have some limitations. For example, the satellite signals may be shielded indoors. Moreover, millions of UEs, e.g., in China, do not have location chips. Cellular network optimization is also complicated and highly costly if using satellite based technologie. Cell identifier (CI) based methods, such as angle of arrival (AOA), time of arrival (TOA), time difference of arrival (TDOA) and MR based methods [1]-[9], utilize CI to indicate the serving cell and the position of an UE. But realizing most of them always need for extra sets or line of sight (LOS) environment [1]-[4]. MR based method could be combined for network optimization due to the available information, such as received signal strength (RSS) and signal-to-noise ratio (SNR).

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 278–285, 2010.

An UE measures signals emitted from the nearby cells periodically, generates MRs, and sends them to its serving cell. In a GSM system, an MR includes received signal level (RXLEV), received signal quality (RXQUAL), discontinuous transmission (DTX), broadcast control channel (BCCH), and base station identity code (BSIC) [12], similar for UMTS [13]. The MR is then forwarded to the base station controller (BSC) and could be intercepted by ABIS interface of GSM or Iub interface of UMTS. On the other hand, engineers of mobile companies have measured configurations of cells, such as longitudes, latitudes, antenna azimuths, BCCHs and BSICs of cells, and stored the information in CCD. As the virtual antenna algorithm which combines all the above information has been suggested by us, we keep on research for further accuracy enhancements.

2 Previous Work

2.1 Field Measurement

To get the practical MRs, we carried out a field measurement in Huangshi with the help of Huangshi Branch of China Mobile Communications Corporation (CMCC). Huangshi is a middle-scale city of China. It covers 227 square kilometers. The day was sunny, and we started most of the calls in the urban district. The user state included statistic, pedestrian and vehicle. Power control modules of the GSM network cells were all opened, and a GPS set was connected to notebook PC for practical position records. The data collection set is the drive test tools introduced in following: The drive test system called TEMS is developed by Ericsson. It includes a phone for generating MRs, 2 data wires with USB and serial interfaces, the supervisory application running on a windows system of the notebook PC. The GPS set BU 353 is produced by Globalsat [14]. The horizontal accuracy position is 10 meters, and the velocity accuracy is 0.1 m/s.

2.2 Enhanced CI-RXLEV Method

There are many MR based methods, such as CI, CI-PLAIN, CI-RXLEV, CI-ANGLE, fingerprints, virtual antenna, etc. [5]-[11]. Fingerprints methods could get higher accuracy than the others in theory. It calibrates signal strength (SS) for building radio map, and locates UE by the database record matches [8], [9]. But construction of the fingerprint database requires great human resource and also the maintenance is costly. So we suggested the enhanced CI-RXLEV (CI-RXLEV-E) method.

CI-RXLEV-E is the improvement of the CI-RXLEV algorithm which considers that network cells spread in a 2-dimension plane, and emission powers of all cell antennas are the same [10]. As the distance between the UE and each cell positively relates to RXLEV, UE position could be estimated by considering RXLEVs of the cells as position weights. During our previous experiments, it was found that the position error size (PES) always grew up with the increase of the distances between serving cell and neighbor cells. So we advanced a threshold of the distances for abandoning the neighbor cells which were far from the serving cell. It’s described as

$$(x,y)=\left(\frac{\sum_{i=1}^{N}\varepsilon_i x_i}{\sum_{i=1}^{N}\varepsilon_i},\frac{\sum_{i=1}^{N}\varepsilon_i y_i}{\sum_{i=1}^{N}\varepsilon_i}\right), \tag{1}$$

$$\varepsilon_i\left(R_i,d_{SN}\right)=\begin{cases}0, & d_{SN}>f_{r-dsn}\\ R_i, & d_{SN}\le f_{r-dsn}\end{cases}, \tag{2}$$

where (x, y) is the UE position and to be solved. (x_i, y_i) is the longitude and latitude of the cell whose identifier number is i. N is the number of all cells contained in a MR, and maximal value of N is 7. R_i is the RXLEV which relates to cell i. d_{SN} is the distance between the serving cell and neighbor cell contained in a MR. $f_{r\text{-}dsn}$ is the threshold, and it's an empirical value.

2.3 Virtual Antenna Method

Virtual antenna method is based on CI-RXLEV-E. It enhances the position accuracy due to the enriched location information, such as combination of multiple MRs, classification of the cells, and cell antenna azimuths. The combination of multiple MRs includes many MRs generated by the same UE in a short time interval. As the maximal number of the cells which belong to the same Node B is three, we classified the cells contained in a combination into three categories: the cell which belongs to a Node B, and there is no other cell contained in the combination belongs to the Node B; the cell which belongs to a Node B, and there is only another cell contained in the combination belongs to the Node B; the cell which belongs to a Node B, and there are only other two cells contained in the combination belong to the Node B.

During our previous experiments, it was found that the cells which belong to the same Node B always were always nearby the UE. So the RXLEV of a virtual antenna is defined for improving the position weight of cells. It's described as

$$\varepsilon_{N_v}=f_{rv-N_v}\sum_{i=1}^{N_v}\varepsilon_i, \tag{3}$$

where ε_{Nv} is the RLXEV of virtual antenna, N_v indicates the category of cell i and could be endowed with 1, 2, or 3. As the second and third category cells are always nearby the UE, $f_{rv\text{-}Nv}$ could be seemed as a coefficient of RXLEV, and (3) should be substituted for (2) to enhance the position weight. As numerator and denominator of (2) both utilize $f_{rv\text{-}Nv}$, $f_{rv\text{-}1}$ could be normalized as 1.

Also the positions of MRs calculated by CI-RXLEV-E could be averaged for more accurate results. It's described as

$$(x,y)=\left(\sum_{j=1}^{K}x_j/K,\sum_{j=1}^{K}y_j/K\right), \tag{4}$$

where K is the number of MRs, (x_j, y_j) is the CP calculated by (1). Our previous experiment proofed that the accuracy (5) is higher than CI-RXLEV-E.

The azimuth of virtual antenna is to indicate the direction angle from the Node B to UE. So the positions calculated by (4) could be improved by azimuth rotation step by step.

3 Accuracy Enhancements

3.1 Relative Definitions

Before introductions of the accuracy enhancements, some relative definitions are explained as below.

GP. It's the position of UE recorded by GPS set. It could be seemed as the practical position.

CP. It's the position of UE calculated by the location methods.

3.2 Recursion of Virtual Antenna Results

As (2) describes, the threshold of distance between the serving cell and the neighbor cell is set for filtering cells. As the GP could be gotten, position of the serving cell is seemed as the GP, and the treatment has achieved good effect. Most of the neighbor cells which are far from the UE are abandoned, and the cells nearby the UE are saved. But there are also some cases that the serving cell is far from the UE. In those cases, neighbor cells which are far from the serving cell but nearby the UE are abandoned, and and some cells which are far from the UE are saved. Also the serving cell should also be filtered.

As the CP calculated by the virtual antenna method has been improved significantly, it could be substitute for position of the serving cell, and the distance between the CP calculated by the virtual antenna method and a neighbor cell could be substitute for the d_{SN} of (2). So the availability of filter could be enhanced. The improvement of (2) is based on the first calculation results of original virtual antenna method, and it's described as

$$\varepsilon_i\left(R_i, d_{VN}\right)=\begin{cases}0, & d_{VN} > f_{r-dsn} \\ R_i, & d_{VN} \le f_{r-dsn}\end{cases}, \tag{5}$$

where d_{VN} is the distance between the CP calculated by virtual antenna method and cell i. R_i is the RXLEV which relates to cell i. cell i could be not only the neighbor cell but also the serving cell. $f_{r\text{-}dsn}$ is the empirical threshold, and should be refitted. (5) is substituted for the second utilization of the virtual antenna method, and the CP should be recalculated.

3.3 The New Parameter of MPPE Criterion

MPPE criterion is utilized for estimating PES calculated by the combination of multiple MRs [11]. As the information of CI-RXLEV-E includes RXLEVs and positions of the cells, statistical parameters of each combination could be extracted from the information, such as the average and standard deviation of RXLEVs, the average and standard deviation of distances between each pair of cells. The distance is always available for estimating as its fluctuation is always bigger than the fluctuation of RXLEV. The distance between each pair of cells is always from several hundreds of meters to thousands meters, but the RXLEV is always from 30 to 63. So as (2) and (4) describe, PES will increase with the growth of average distance. RXLEV relates to propagation path non-linearly, average and deviation of it should also be utilized. As the average distance between each pair of cells always increase with growth of the cell number, MPPE criterion is described as

$$score = \mu(d)^{f_{score-ad}} \times D(\varepsilon)^{f_{score-dr}} \times \mu(\varepsilon)^{f_{score-ar}} \times N^{f_{score-n}} , \tag{6}$$

where score is the result to estimate PES. The larger score, the larger PES might be. d is the distance between each pair of virtual antennas in a combination, D is the function to solve standard deviation, μ is the function to solve average, ε is the RXLEV of virtual antenna i. $f_{score-ad}$, $f_{score-dr}$, $f_{score-ar}$ and $f_{score-n}$ are index coefficients which have been fitted. Also a low limit value of 0.5 is set for avoiding D(ε) approaching 0.

As MPPE score reflects the PES calculated by (4), rotation expression of virtual antenna method utilizes it for indicating rotation degree [14]. The rotation expression is

$$\vartheta_i = f_{\vartheta-Nv} d_{i-c}{}^{f_{\vartheta-d}} \overline{\varepsilon}_i{}^{f_{\vartheta-r}} score^{f_{\vartheta-score}} clip\left(\theta_{c-i}, \Phi_{Nv-i}\right) , \tag{7}$$

where ϑ_i is the rotation degree which reflects the rotation step size, ε_i, d_{i-c}, and clip are all factors to indicate the rotation degree. Score is calculated by the MPPE expression. $f_{\vartheta-Nv}$, $f_{\vartheta-d}$, $f_{\vartheta-r}$ and $f_{\vartheta-score}$ are empirical index coefficients.

It could be found that if score approaches 0, rotation degree will approach 0 too. So we consider importing a new parameter of the score expression for avoiding score approaching 0. The improved MPPE criterion is

$$score = f_{score} + \mu(d)^{f_{score-ad}} \times D(\varepsilon)^{f_{score-dr}} \times \mu(\varepsilon)^{f_{score-ar}} \times N^{f_{score-n}} , \tag{8}$$

where f_{score} is the new parameter. f_{score}, $f_{score-ad}$, $f_{score-dr}$, $f_{score-ar}$ and $f_{score-n}$ should all be refitted.

4 Verification

4.1 Fitness

We applied over 14000 MRs generated by 140 calls for fitness. The least square method is utilized for fitting empirical coefficients of (7), and described as

$$M = \min\left\{\sum_{i=1}^{n} e^2\right\} , \tag{9}$$

where n is the number of all samples. e denotes the difference between the PES calculated by (4) and the MPPE score. Table 1 shows the empirical values of the parameters.

Table 1.

Parameter	Range	Minimal Step Size	Fitness Value
f_{score}	[0, 100]	5	70
$f_{score-ad}$	[-3, 3]	0.05	1.25
$f_{score-dr}$	[-3, 3]	0.05	-1.80
$f_{score-ar}$	[-3, 3]	0.05	0.95
$f_{score-n}$	[-3, 3]	0.05	-0.80

4.2 Accuracies Comparison

New enhancements of the virtual antenna method are all utilized, and we select 10000 MRs which haven't been fitted for accuracy verification. PESs calculated by the new enhancements and CI-RXLEV-E are compared as Fig. 1.

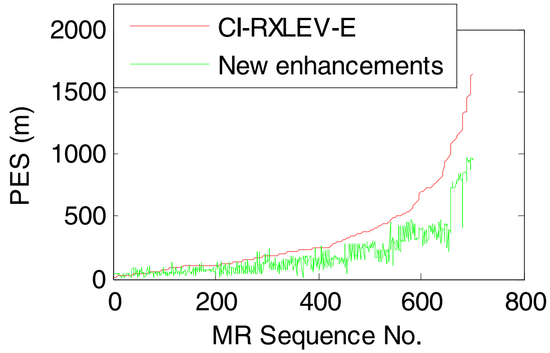


Fig. 1. The PESs of the two methods are ordered ascendant by the PESs of CI-RXLEV-E. It could be found that our new enhancements always perform better than the previous suggestion.

Also the statistics of PESs calculated by many location methods are compared.

Table 2.

Location method	PES (meters)	
	50%	90%
New Enhancements	175	408
Virtual Antenna	183	428
CI	347	940
CI-RXLEV-E	215	767
Fingerprints (Down town)	94	291
Fingerprints (Residential)	277	984

As Table 2 shows, the data of fingerprints is cited from [7]. It could be found that the new enhancement performs better than the original virtual antenna method, and it's close to the fingerprints. The average PES of new enhancements is 193.31 meters.

5 Conclusion

In this paper, we suggest some enhancements of the virtual antenna method for UE location. The virtual antenna method utilizes antenna azimuths and some empirical attributes of the cells to enhance position accuracy. Based on the positions calculated by the virtual antenna method, the cells which are near by the serving cell but far from the UE would be abandoned more accurately, and the serving cell would be filtered also. A new parameter is imported for the MPPE criterion, and so the MPPE score is always larger than a positive certain value. Rotation of the virtual antenna method is more available. The experiment results shows that though the amplitude of accuracy enhancement is limited, it provides a new basic point of the further research for MR based methods.

References

1. Wann, C.D., Lin, H.Y.: Hybrid TOA/AOA estimation error test and non-line of sight identification in wireless location. Wirel. Commun. & Mobile Computing 9(6), 859–873 (2009)
2. Chen, C.S., Su, S.L., Huang, Y.F.: Hybrid TOA/AOA Geometrical Positioning Schemes for Mobile Location. IEEE Trans. Commun. E92B(2), 396–402 (2009)
3. Mazuelas, S., Lago, F.: Ranking of TOA Measurements Based on the Estimate of the NLOS Propagation Contribution in a Wireless Location System. In: Wireless Personal Commun., Euro-Par 2010, vol. 53(1), pp. 35–52. Springer, Heidelberg (2010)
4. Li, W.C., Wei, P., Xiao, X.C.: A robust TDOA-based location method and its performance analysis. Sci. Chin. Series F-Info. Sci. 52(5), 876–882 (2009)
5. ETSI.: Digital telecommunications system (Phase 2+); Location Services (LCS); Functional description; Stage 2 (3GPP TS 03.71 version 8.9.0 Release (1999)
6. Weckstrm, M., Spirito, M., Ruutuu, V.: Mobile station location. In: GSM, GPRS and EDGE Performance, New York, pp. 119–141 (2002)
7. Chen, M., Sohn, T.: Practical metropolitan-scale positioning for GSM phones. In: Dourish, P., Friday, A. (eds.) UbiComp 2006. LNCS, vol. 4206, pp. 225–242. Springer, Heidelberg (2006)
8. Fang, S.H., Lin, T.N.: A Dynamic System Approach for Radio Location Fingerprinting in Wireless Local Area Networks. IEEE Trans. Commun. 58, 1020–1025 (2010)
9. Bshara, M., Orguner, U., Gustafsson, F.: Fingerprinting Localization in Wireless Networks Based on Received-Signal-Strength Measurements: A Case Study on WiMAX Networks. IEEE Trans. Veh. Tech. 59(1), 283–294 (2010)
10. Fu, T., Huang, B.X.: How About the Real Capacity of Cell-ID Based Algorithms for LB. In: Internet Conference on E-Business and E-Government (ICEE), IEEE Press, Guangzhou (2010)

11. Fu, T., Huang, B.X., Mo, Y.J.: The MPPE Criterion for Estimating Position Error Size of CI based Method. In: Internet Conference on Wireless Commun., Network and Mobile Computing (WICOM), IEEE Press, Chengdu (2010)
12. ETSI.: Digital cellular telecommunications system[S] (Phase 2+); radio subsystem synchronization (GSM 05.10 Ver.7.3.0 Rel (1998)
13. 3GPP.: Universal mobile telecommunications system (UMTS); User equipment (UE) positioning in universal terrestrial radio access network (UTRAN) [S].Stage 2 (3GPP TS 25.305 Ver.7.1.0 Rel.7)
14. Globalset, http://www.globalsat.com.tw/eng/product_detail_00000044.htm

Research on the Production Scheduling Management System Based on SOA

Baoan Li

Computer School
Beijing Information Science and Technology University, Beijing, China
liba@bistu.edu.cn

Abstract. The production scheduling management is an important job which directs the whole production activities in a refinery enterprise. It is an important measure to build the dynamic scheduling management information system to improve the work efficiency, promote the production to standardize, and increase the enterprise competition ability. But how to connect the new technologies with the value chains of enterprise is more concerned by the enterprise users. After analyzing the service requirements in the refinery enterprise in depth, it develops the component libraries and then gives a solution to build the scheduling management system based on SOA and component oriented technologies. Otherwise the development of enterprise software, the program method of SOA, the interrelated SOA standards and component oriented technologies are presented in this paper.

Keywords: Component Oriented; SOA(Service Oriented Architecture); Scheduling Management; ERP(Enterprise Resources Planning).

1 Introduction

The development of enterprise software is a very important field in many countries. As an important methodology in recent years, SOA (Service Oriented Architecture) brings more and more attention. With the change of the users' requirements towards the way fitting every-change environment, SOA and component oriented technologies turn to the core technologies to establish ERP (Enterprise Resources Planning), CIMS(Computer Integrated Manufacture System), EERP(End-to-End Resources Planning), and other large information systems. And then the question about the program method of SOA and integration of the services becomes more and more urgent and important. Otherwise, the most concern question coming from enterprise is how to connect these technologies to the enterprise different value chains so that to adapt to the requirements. All these questions should be considered seriously when we design the enterprise information systems.

2 Research on Enterprise Application Software

The development of enterprise application software shows as Fig. 1.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 286–294, 2010.

The development of enterprise application software, such as ERP and EERP can be described as changing from the centre of resources towards the centre of relation. On the progress of the Enterprise Degree Software, we has developed more software from MRP (Material Requirements Resource Planning), MRPII (Manufacturing Resource Planning), ERP(Enterprise Resource Planning), ERPII (Enterprise Resources Planning and Co-Business) towards EERP. Emphatically, the relations amongst them are development or including, not replacement or negation.

In recently years, SOA becomes hotter and hotter [1]. SOA and component oriented technologies offer the new way to accomplish large systems such as ERP/EERP on the more scope relation platform.

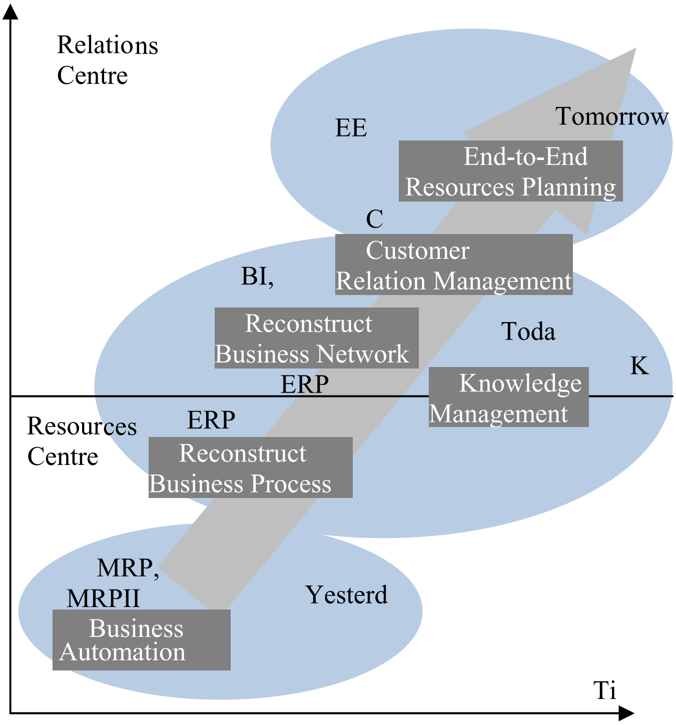


Fig. 1. The Development of Enterprise Application Software

3 Research on SOA and Compentent Oriented

A The development of SOA

The development processes are divided into three phases:

Firstly, SOA focuses on the integration in enterprise and resolves the One to One relations. It starts from 2003 year.

Secondly, SOA focuses on the value chains between the credible associate enterprises and resolves the One to Many relations. It begins from 2007 year.

Thirdly, SOA Focus on finding new associate and new services. It resolves the Many to Many relations. However, the start year of this phase cannot be confirmed. It depends on the advanced research and application about SOA and other new technologies [2].

In SOA, services are wrapped as loosely coupled reusable web services. However, at implementation time, there is no way to loosely couple a service or any other interaction between systems.

The systems must have some common understanding to conduct an interaction. Instead, to achieve the benefits of loose coupling, consideration should be given to how to couple or decouple various aspects of service interactions, such as the platform and language in which services are implemented, the communication protocols used to invoke services, and the data formats used to exchange input and output data between service consumers and providers. ESB (Enterprise Service Bus) can carry the services out between requesters and providers.

In Fig. 2, ESB is depicted as a logical component in SOA. It acts as the mediator between the service consumers and services providers. The service providers and service consumers never interact directly. The ESB provides services to resolve differences in protocol and format, and decouple the service consumer from the service provider [3].

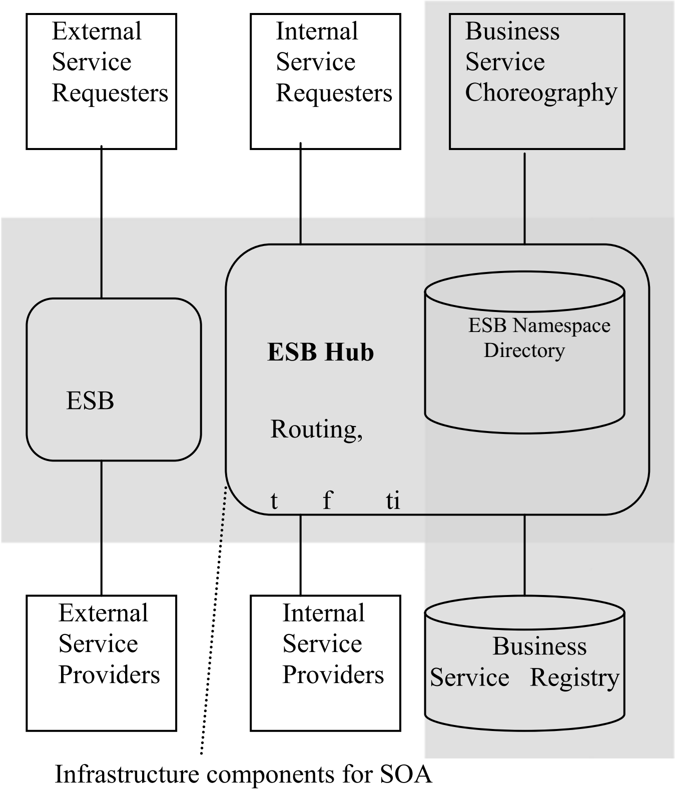

External Service Requesters
Internal Service Requesters
Business Service Choreography
ESB
ESB Hub
Routing,
t f ti
ESB Namespace Directory
External Service Providers
Internal Service Providers
Business Service Registry
Infrastructure components for SOA

Fig. 2. ESB and SOA

In order to support service discovery, selection and composition for process optimization, quality of services (QoS) and service level agreement (SLA) are required to fully meet business goal.

B The program method of SOA

The component oriented technique is an effective and feasible method to actualize SOA. This is also the demand of the roadmap of SOA in China [4]. The components are assembly and binding according to the business process though ESB. They should support the SCA (Service component Architecture)/SDO (Service Data Object) standards so that to be reused in the ever-changed environments [5].

At the present time, the famous standard organizations and industry alliances have issued a lot of standards or criterions. Hereinto SCA/SDO is the important standards about component oriented technologies [6]. The Component Oriented Programming shows as Fig. 3.

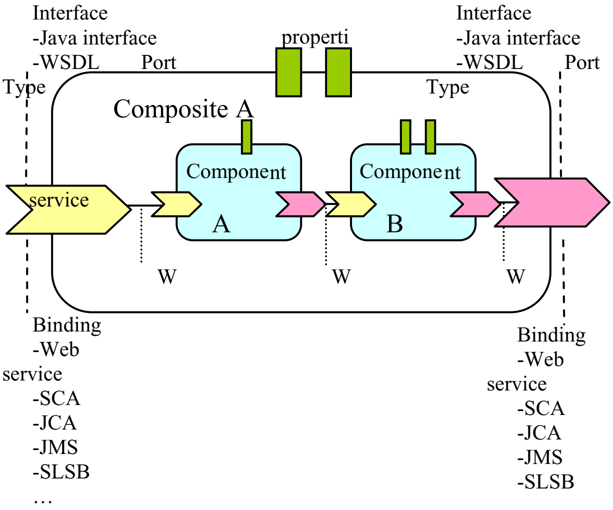


Fig. 3. The Component Oriented Programming

4 A Refinery Production Scheduling System Based on SOA

A Analyzing the profit points of the production scheduling business

The profit points of the production scheduling business in refinery shows as Fig. 4.

The scheduling core business is on the basis of production data collection, analysis, and dynamic instruction, control of production.

Scheduling activity in production logistics activities for each link has played an important role. It is a vital link in the benefit of the enterprise.

The critical point of scheduling management is: monitoring of energy consumption in the production process; keep the material balance; timely adjustment production directive according to the product quality; control of raw materials and products well provisioned work; and to maximize the potential of dispatchers, makes its scheduling level to maximize.

The key points of the efficiency of the scheduling work must be gotten so that to improve schedule levels, to improve product quality, to control product yield, and can arrive at the aim of energy saving, optimize production, maximize profits.

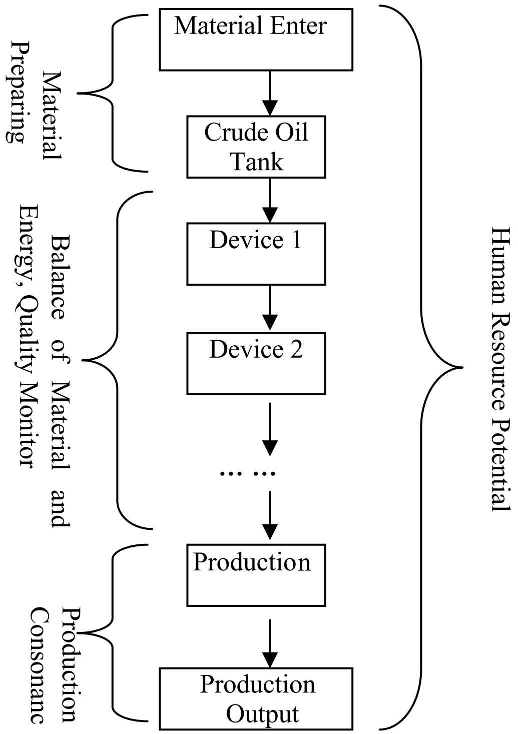


Fig. 4. The profit points of the production scheduling business

B Building the service component libraries

It is important to build the adapted service component libraries to realize the enterprise information systems with component oriented method. Here, we show an example of refineries. The component libraries in refineries' ERP are constructed based on the business value chains. They can be divided into two classes, one is commonly foundation component library and another is professional component library for the special application.

The foundation component library offers the common services, such as the basic calculating components, the basic operation components, the basic application components, workflow components, page label components and especially the ERP Web service functions based on semantic SOA to speed up the realizing of enterprise resource planning and end–to-end resource planning. These services can be gained from the third SOA middle software provider.

On the other hand, the special professional component library provides the professional services to be used more frequently in the special field. The professional services are necessary and important to accelerate the development of the large systems in the special application field. They are nearly relevant with the special value activities.

The components are assembly and binding according to the business process though ESB. Fig. 5 shows the component libraries for the refineries' information systems.

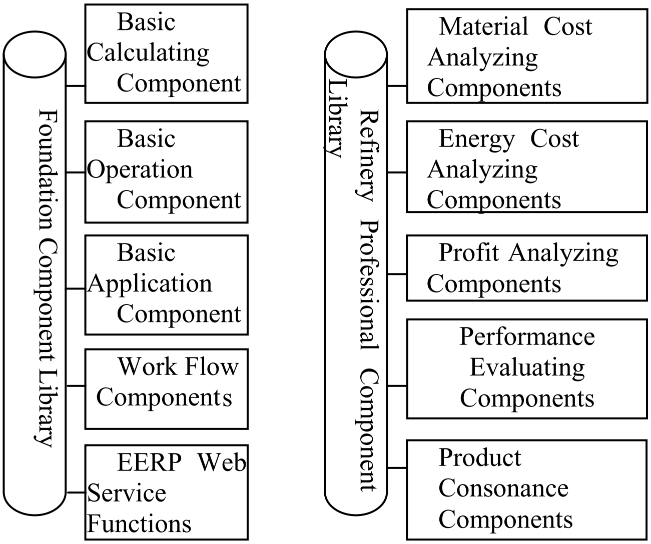

Foundation Component Library
Basic Calculating Component
Basic Operation Component
Basic Application Component
Work Flow Components
EERP Web Service Functions
Refinery Professional Component Library
Material Cost Analyzing Components
Energy Cost Analyzing Components
Profit Analyzing Components
Performance Evaluating Components
Product Consonance Components

Fig. 5. Component libraries for the refineries

C The information flow on scheduling business

Take advantage of production management information system, the petrochemical enterprises scheduling work-related data, analysis, storage, statistics, publications, and other various management links organically together. Systems to enterprise scheduling business as the core, in accordance with the information flow in a production run of logistics processes, design their systems function modules, enabling enterprises to the logistics base in each production unit for real-time monitoring, scheduling, reached the flexible, real-time regulation to optimize production purposes. The information flow of the production scheduling business shows in Fig. 6.

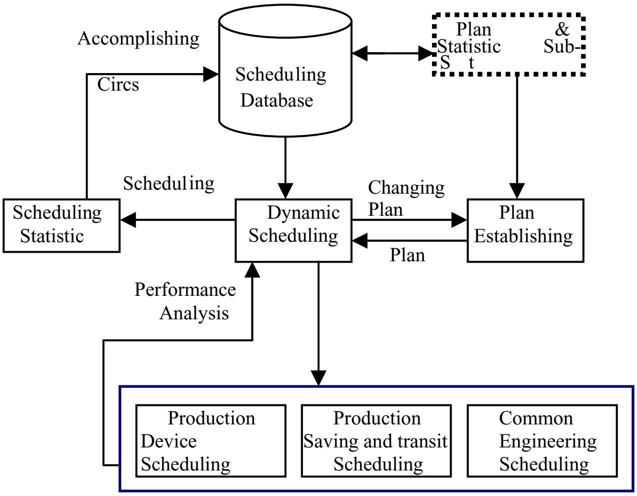


Fig. 6. The information flow of the production scheduling business

D The solution of production scheduling management system

The material balance, consonance plan, emergency management, production cost management are the core modules of production management system. The production data coming from each measurement point in DCS, PLC, and DDZ digital instruments are inputted into the scheduling system in the Real-time foundation database. The data information from the Real-time foundation database can be

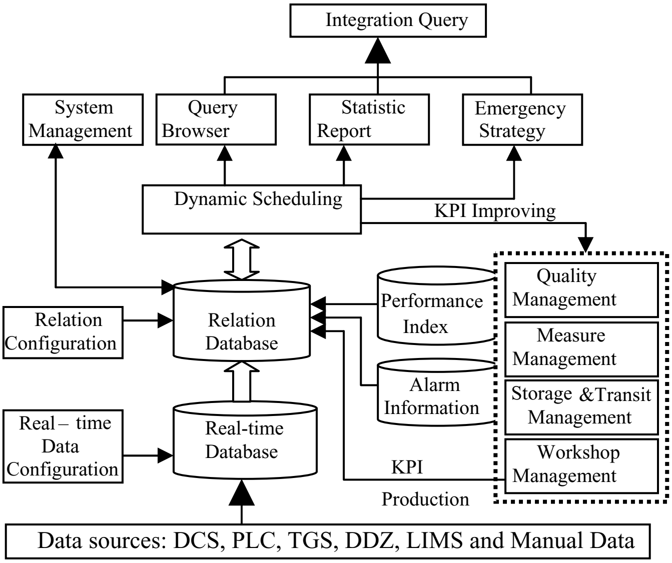


Fig. 7. The struction of production scheduling system

directly used in the dynamic scheduling management module to handle a variety of business; also available to extract data from a relational database. The specific forms of schedule data reports, documents and information can be gotten through the relation configuration.

The related system interfaces with other system have been predesigned, and the component libraries such as foundation component library and refinery professional component library have been used. It speeds up the realizing of enterprise management system and makes it easier for users in system combination and system expansion. It improves system configuration flexibility. Adhering to the overall planning, step-by-step implementation of principles, and the methodology based on SOA and component oriented technology, it bring the enterprise's information work to save money and time, maximize to meet user's requirements, to reduce costs, optimize production and increase their work efficiency.

Fig. 7 shows the solution of production scheduling management system.

5 Conclusion

The good information system can enhance competitiveness for enterprise, increase production and productivity. It is the key about whether the huge investments about the information system for enterprise can reap the rewards. It is important to pay more attention on the comprehensive and practical analysis about various activities and the corresponding value chains (such as production, supply and marketing value chain, material consumption and energy consumption value chain, equipment maintenance value chain, human resources performance management value chain, etc.) [7].

SOA, component oriented technology and SSOA (Semantic SOA) are the effective methods [8] to implement large information systems such as ERP / EERP systems. But these IT technologies must connect with the actual requirements from the specific fields. It is important to develop the service components, and they must be developed according to the value chains in the specific fields. It is not enough and may be very difficult to building the large information systems just using the middle common SOA products. With the research and application about SOA, component oriented technology, SSOA and other Web service technologies go to deeper and deeper, and the development of large information systems will be easier and more effective.

Acknowledgment

This work was supported by Funding Project for Academic Human Resources Development in Institutions of Higher Learning under the Jurisdiction of Beijing Municipality (Grant No. PHR201007131). And it was also supported by Education Reform Project of Beijing Information Science and Technology University (Grant No.2009JG24).

We also thank the experts and the organizations that appeared in the paper references. Their achievements and experiences are very helpful and outstanding.

References

1. Joe, M.: Survey: SOA isn't Just Surviving, It's Thriving. In: ebizQ SOA in Action Blog, March 27, pp. 1–2 (2008)
2. Li, B., Li, M.: Research and Design on the Refinery ERP and EERP Based on SOA and the Component Oriented Technology. In: 2009 International Conference on Networking and Digital Society(IEEE ICNDS 2009), pp. 85–88 (May 2009)
3. Patterns.: SOA Design Using WebSphere Message Broker and WebSphere ESB, IBM.com/Redbooks (2007)
4. IDC, China road map of SOA. IDC White Paper (May 2007), `http://gocom.primeton.com/special/soabook/soa.php`
5. SCA_AssemblyModel v1.0, `http://www.osoa.org/`
6. Service Data Objects For Java Specification, `http://www.osoa.org/`
7. Li, B., Zhou, W., He, Y.: To design Component Oriented ERP System Based on Analyzing Enterprise Value-chains. Computer Engineering and Design 29(15), 3927–3928 (2008)
8. Matthews, B.: JISC Semantic Web Technologies (November 2008), `http://www.cs.umbc.edufininpaperspolicy03.pdf`

A Requirement-Driven Approach to Enterprise Application Development

Jinan Duan[1], Qunxiong Zhu[1], and Zhong Guan[2]

[1] Department of Information Science and Technology
Beijing University of Chemical Technology
Beijing, China
jarenduan@gmail.com,
zhuqx@mail.buct.edu.cn
[2] Department of Publicity and United Front Work
Liaoning Shihua University, Fushun
Liaoning, China
guanzhong1978227@163.com

Abstract. The requirements changes are the root causes of evolution of enterprise applications. How to effectively develop enterprise application with the frequently changing requirements is still a challenge to software engineering. The two main aspects are how to capture requirements changes and then how to reflect them to the applications. Use cases and refactoring are excellent tools to capture functional requirements and to change object-oriented software gradually. This paper presents a requirement-driven approach to enterprise application development. The approach uses refined use cases to capture the requirements and to build domain models, controller logics and views. It transforms requirement changes into the refactorings of refined use cases, thus it can propagate the modification to the application. With rapidly continuous iterations, this approach tries to give a solution to the problems of enterprise applications development.

Keywords: enterprise application, requirement-driven, refined use cases, refactoring.

1 Introduction

Enterprise applications have been already deployed and applied in many industries or companies in the modern world. These applications are built for fulfilling the requirements from different users. However, the requirements are usually constantly changing during the enterprise application development process or even the whole life-cycle[1]. The changes often induce many problems to the development, such as rebuilding domain models, continuously refactoring, chains of modifications of coupled components[2] and unexpected iterations, which bring developers and users lots of costs.

Traditional ways like UML would suggest developers start developing application after obtaining elaborate requirements analyze results. It is very useful to develop

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 295–302, 2010.

large software with such fine design. But unfortunately, in many cases, especially for the modern enterprise applications whose requirements evolves fast, requirements changes are hard to precisely estimate and cannot be obtained before development. Thus since the early 90s, refactoring methodology[3] and agile process[4] became more and more popular and have given us much guidance in practice. Although those methods still cannot solve the problem perfectly, they give us many meaningful tools to explore further. For example, use case is very useful to capture almost all of the functional requirements from users[5], and refactoring can be used to gradually modify object-oriented applications in different levels[6].

How to develop enterprise application under the constantly changing requirement? This question contains two aspects from our point of view:

1. How to capture primitive requirements, and then transform them into specific requirement and specification such as domain models, business logics and the views of applications.
2. How to effectively propagate the changes to the application.

Aiming at the characteristics of enterprise application development, we propose a requirement-driven approach. The approach uses refined use cases(RUCs)[7] to capture requirements and build domain models, controllers and view models. It transforms the requirement changes into a series of refactorings of RUCs, and then maps them to the models of application to modify the codes. During the continuously iterations, it helps us to complete the development process.

The rest of this paper is organized as follows. Section 2 discusses the approach in general. Section 3 introduces some details of how to refine requirement into RUCs. Section 4 discusses a few of issues about modeling application. Section 5 discusses the refactoring of RUCs, changes propagation and iterations. Section 6 introduces some related works of other researchers. Section 7 gives the conclusions.

2 An Approach Overview

The approach presented is driven by requirements, that is, all of the developing activities start from requirements or user feedbacks. In the beginning of development process, analysts usually obtain primitive requirements by using drafts or interactive questioning method. It often forms requirement text like the text use case. Analysts need to refine them into one or more RUCs.

RUC based on use case is an approach for formal requirement specifications. It is very useful to describe the interactions between the environment and the computer system. To get RUCs, we use refined use case description language (RUCDL), which we'll discuss in the next section. Once we get the RUCs, developers can build domain models, controller and view models by parsing RUCs in accordance with established modeling rules. And then we finish the code implementation and post the application to user to test it.

In most cases, testing result will bring feedbacks, which may include program bugs, interface modifications or rearrangement of work flow. It's also possible that users will raise new requirements after the test. In some cases, test will fail because of the misunderstandings between requirement analysts and users in the previous phrase.

This could be due to many reasons such as asymmetric information between two sides.

In either case, users will propose the requirements changes. Requirement analysts have to keep communicating with users to get latest information. Developers have to transform these modifications into the refactorings of RUCs so that it can do small changes to application while observing the behavior of the other parts of application which have already satisfied requirements. Developers then map the refactorings of RUCs to the models of application, and change the codes, rebuild application, submit to users, finally leading us to the next iterative process.

Through repeated feedbacks and iterations, we move forward a small step based on the last version each time. The refactoring we executed can maintain the consistency of application model, so that we can keep producing reliable application that fulfills the current requirements, until releasing out the final version. Fig.1 shows a typical process of using this approach.

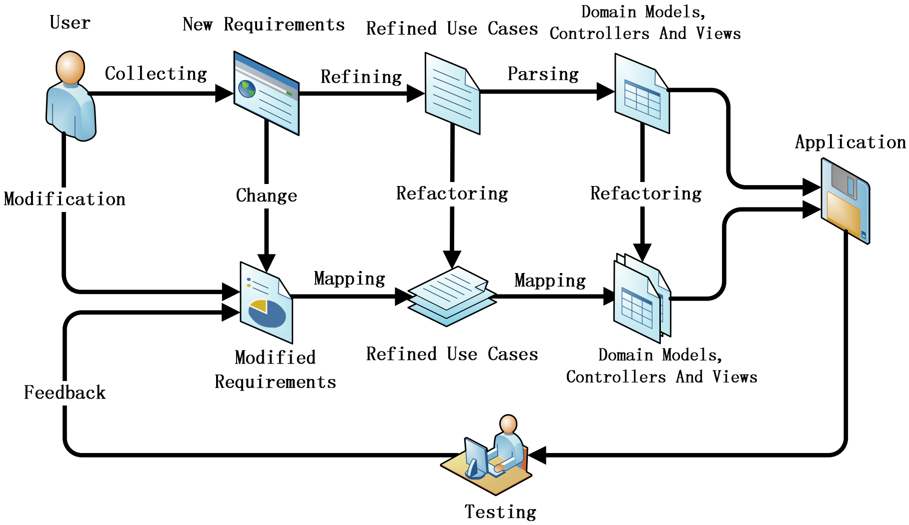


Fig. 1. A typical process of requirement-driven development

3 Refining Requirements

Use cases are often used to capture requirement and build models when moving from requirements analysis to design[8]. However, it is not easy to elicit object-oriented models from use cases written in plain text. First of all, many use cases using in industries are use case diagrams, which do not include the specific functions deep inside. Secondly, text use cases are usually written in natural language without fixed format, therefore it's hard to parse by computer. And the biggest disadvantage is that it is difficult to keep the semantics of requirement changes, that is to say, we've lost information of modification activities during the changing process, only storing the final state of requirements. Due to the limitations of use case, the approach adopts a formal method called refined use cases which is written in RUCDL, and tries to keep the semantics of changes through refactoring of refined use case.

In general, refining requirement to RUCs involves four steps:

1. Checking the consistency of requirements. New requirements gathered from users or feedbacks may conflict with the current requirements and also bring potential logic errors. It is usually represented as the semantics confliction of RUCs. For example, users could use the same identifier to refer different entities. Thus the analysts need to check the consistency before they modify RUCs, which includes verifying the semantics and duplication of existing identifiers, detecting type confliction introduced by assignment or view settings, eliminating the logic errors in work flow, etc.

2. Deciding the preconditions and the constraints. Analysts often have to decide several conditions like user authentication to satisfy when entering a RUC. There are also constraints to consider about the data of entities.

3. Building the business work flow. Analysts need to describe the operations of users and the respond of system. They decompose the requirements into several simple short sentences rather than using long and complicated sentences, so that they can be easily expressed by description language.

4. Describing views. Modern enterprise applications often require convenient UIs for interacting. RUC support views by using a View section in the RUC document. For the work flow only focuses on data flow and data relationships, analysts have to separate the UI descriptions from the work flow. Analysts need fill the section according to the UI requirements.

Being more like declarative programming, RUCDL only describes what user does, system responds and the contexts. The core BNFs of RUCDL shows as follow:

```
  Case → id { Pre FLow View}
  Pre → preconditon { As }
  Flow → flow { As }
  View → view { Sets }
  Acts → Act Acts | λ
  Act → id { Acts } OrAct| fill {Acts} post {Acts} OrA |
          Subj Verb Obj Comp ; OrA |While ( Exp ) { As }
|
          if exp { As } else { As } OrA | if exp { As }
OrA
  OrAct→ or Act OrAct | λ
  Subj → user|sys
  Verb → is | select | input | choose | add | print ...
  Obj → ObjectID OtherObj | Set OtherObj
  OtherObj → , Obj | λ
  Prep →in | to | as | from | by
  Comp →Prep Obj | λ
```

RUCDL is a project maintained by BUCT Intelligent Engineering Laboratory, which now is still evolving under development. There are still some very complicated requirements cannot be transformed to refined user cases by RUCDL at present. With more and more features being added, it can support even more and much complicated requirements in the future.

4 Building Domain Models, Controller and View Models

From practical point of view, there are relatively mature patterns of modeling modern enterprise application built on popular technologies, such as Java EE or .NET platform. An enterprise application usually is divided into a few logical layers, such as entity persistence layer, business logic layer and UI layer. Developers tend to use design patterns, libraries or frameworks rather than building application from scratches. Developers can easily finish many common tasks with these conventions of enterprise application development.

On the other hand, all RUCs represent the refined functional requirements of application, which contain most of knowledge of how to build application. By parsing RUCs, we can build domain models, controller and view models[7].

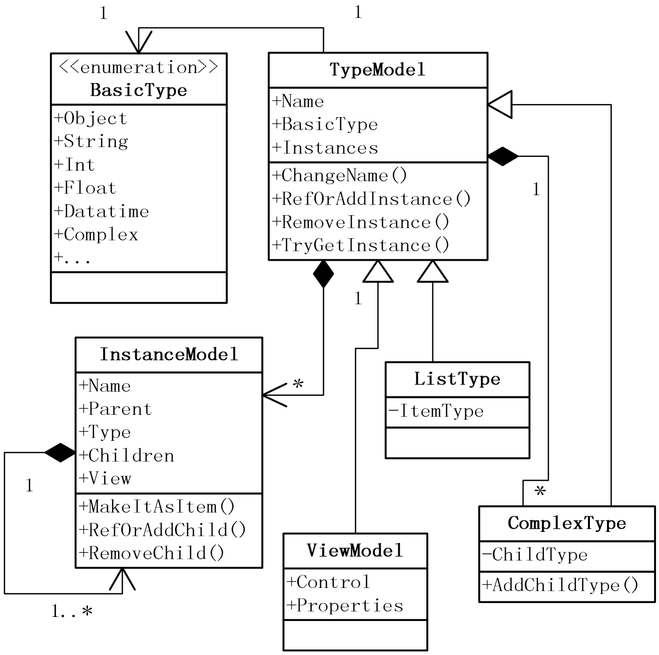


Fig. 2. Core meta-models of instance-type system

RUCs are written for specific case, thus everything in it is just an instance. To get type information, we use an instance-type system to build those models. Fig.2 shows the core meta-model of instance-type system. An instance refers to an entity or a data member within a RUC. RUCDL distinguishes different instances by using alias such as Product-A and Product-B. Type of instance can be inferred by the semantics of activities or by the assignment operations in the workflow and the views. A controller is actually a complex type corresponding to a group of closely related functions. We can map a RUC to a controller, including the data members and control structures like loops or branches.

View models are the visualization of background data, thus a view model is associated to an instance model. A view model contains several parts: layout, styles and behaviors. A good view model usually requires professional art designers.

5 Propagating Changes and Iterating

Refactoring methodology for object-oriented programming provides a set of specific methods of reconstruction. Refactorings check enabling conditions to ensure that a change can be made safely[6]. The ability of changing object-oriented applications makes it become a powerful tool for software evolution.

This approach transforms requirement changes into refactoring of RUCs, so that it can reserve the semantics of changes during the process and keep the consistency of RUCs. The refactorings are performed on the meta-models of RUCs. Fig.3 shows the simplified meta-models of RUCs.

Thus when the requirement changes, developer don’t modify RUCs by hands, but perform corresponding refactoring to the related RUCs, such as AddCase, RemoveCase, RenameObject, ChangeAct, AddAct, RemoveAct, ChangeActStructure,

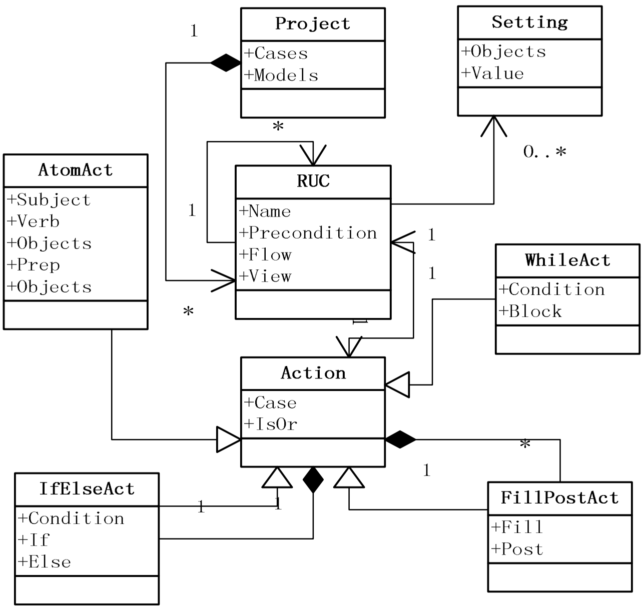


Fig. 3. Simplified meta-model of RUCs

ChangeSubject, ChangeVerb, ChangeObject, AddView, RemoveView, AddSetting, RemoveSetting, AddPrecondtion, RemovePrecondition, etc.

And to propagate the changes, developers then map the RUC refacorings to the model refactorings. Main model refactorings include FoundInstance, FoundType, RemoveInstanceRef, AddPropertyToView, RemovePropertyToView, etc.

The models generated by the approach are platform-independent, thus developers can compose all the models into executable codes for specific platform such as server-side application like ASP.NET or Jave EE projects. After that, developers immediately submit the application to final users to test and evaluate through traditional compilation and deployment process. As we discussed above, feedbacks and new coming requirements will keep bringing changes to existing requirements, which leading us to the next iteration.

6 Related Work

Enterprise application development is so complex and significant that many researchers have been working on it. The work most closely related to the content of this paper is as follows:

6.1 Executable Use Cases(EUCS)

The concept of EUCs proposed by Jens B. Jørgensen et al [9] is very similar to RUCs', especially about how to refine the requirements into the specifications. The main difference is that EUCs is a way to formal requirements modeling, and RUCs are designed to modeling specifications and refactoring for enterprise applications. Another distinct difference between EUCs and RUCs is their formal method. EUCs are using coloured Petri nets while RUCs are using RUCDL.

6.2 Cascaded Refactoring

Lugang Xu[10] presented a cascaded refactoring methodology for framework development in his doctoral thesis. Although the method focused on framework and software product line development, which differs from our application field, it still gave us very useful idea that we can map refactorings between different phrases of application development process so that we can propagate changes to the code through RUCs and application models.

6.3 Requirements-Driven Development Methodology

A requirements-driven development methodology is proposed by Jaelson Castro et al[11] arguing that the modern Information systems have suffered from an impedance mismatch. It and our approach have a lot of in common about requirement should come back to the central of development. The most difference between them is the former is based on agent-oriented, and ours is based on object-oriented paradigm.

7 Conclusions

The requirement-driven approach presented in this paper provides strong help for most of enterprise application, especially for the small and medium enterprise application which has huge market potential. The approach doesn't need cumbersome requirement analysis. It is a light-weight enterprise application development approach. It is competent for RAD and prototype development. By introducing related tool, it can automatically complete much heavy repetitive work such as RUCs management, models refactoring and code generation, which greatly enhance the development efficiency and customer satisfaction. But for some very complicated scenarios, the approach still have a few shortages, for example, the ability of expressing requirements of RUCDL, and the redundancy issue of RUCs. Those are the directions of future researches.

References

1. Godfrey, M.W., German, D.M.: The Past, Present, and Future of Software Evolution. In: International Conference in Software Maintenance (ICSM) Frontiers of Software Maintenance. IEEE Press, Los Alamitos (2008)
2. Visser, J.: Coupled Transformation of Schemas, Documents, Queries, and Constraints. Electronic Notes in Theoretical Computer Science, vol. 200, pp. 3–23 (2008)
3. Opdyke, W.F.: Refactoring Object-Oriented Frameworks. Doctoral Thesis. University of Illinois (1992)
4. Abrahamsson, P.: Agile Software Development Methods: Review and Analysis. VTT Publications, Finland (2002)
5. Cockburn, A.: Writing Effective Use-Cases. Addison-Wesley, Reading (2002)
6. Tokuda, L., Batory, D.: Evolving Object-Oriented Designs with Refactorings. Automated Software Engineering 8, 89–120 (2001)
7. Duan, J.: An Approach for Modeling Business Application Using Refined Use Case. In: International Colloquium on Computing, Communication, Control, and Management, pp. 404–408. IEEE Press, Wuhan (2009)
8. Overmyer, S., Lavoie, B., Rambow, O.: Conceptual Modeling through Liguistic Analysis Using LIDA. In: 23rd International Conference on Software Engineering, pp. 401–410. IEEE Computer Society Press, Toronto (2001)
9. Jørgensen, J.B., Tjell, S., Fernandes, J.M.: Formal Requirements Modeling with Executable Use Cases and Coloured Petri Nets. Innovations Syst. Softw. Eng. 5, 13–25 (2009)
10. Xu, L.: Cascaded Refactoring for Framework Development and Evolution. Doctoral Thesis. Concordia University (2006)
11. Castro, J., Kolp, M., Mylopoulos, J.: Towards Requirements-Driven Information Systems. Engineering Information Systems 27, 365–389 (2002)

ESA: An Efficient and Stable Approach to Querying Reverse *k*-Nearest-Neighbor of Moving Objects[*]

Dunlu Peng, Wenming Long, Ting Huang, and Huan Huo

School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai 200093, China
dunlu.peng@gmail, wilma_long@163.com,
lost_august@yahoo.cn, huoh@usst.edu.cn

Abstract. In this work, we study how to improve the efficiency and stability of querying reverse k-nearest-neighbor (R*k*NN) for moving objects. An approach named as ESA is presented in this paper. Different from the existing approaches, ESA selects *k* objects as *pruning reference objects* for each time of pruning. In this way, its greatly improves the query efficiency. ESA also reduces the communication cost and enhances the stability of the server by adaptively adjusting the objects' *safe regions*. Experimental results verify the performance of our proposed approach.

Keywords: Moving objects, adaptive safe region, R*k*NN, *k*-Binding.

1 Introduction

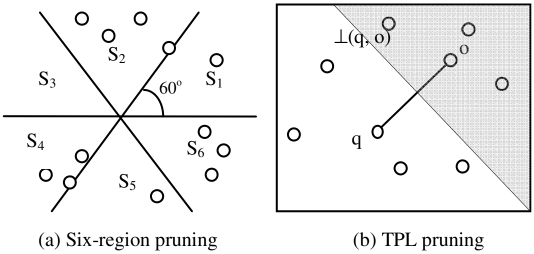
With the extensive use of Location Based Service (LBS), query based on location is playing a quite important role in our daily life. Querying Reverse *k*-Nearest-Neighbor (R*k*NN) is one of the important issues of these queries.

So far, researchers have done much work on RNN and R*k*NN queries. Konrn et al. [1] are the pioneers of RNN query; their approaches are to draw a nearest neighbor cycle for each query object and check whether the query point *q* is on the cycle. If yes, the query object is a member of the query results. Stanoi et al [2] presented six-regions pruning approach for RNN query. Fig.1 (a) shows us this approach. It evenly divided the two-dimensional space into six regions (S_1 to S_6), so there is no more than one RNN in each region. The most likely point is the nearest neighbor, so we just need to pick out the nearest neighbors in each region to verify. It improved the efficiency of RNN query by dividing the RNN query into two phases: filtering phase and verifying phase. Many approaches were proposed based on the two-phase computation [3, 4]. After Stanoi et al. proposed the six-regions pruning, Tao et al. [5] proposed TPL

[*] This work is support by Municipal Nature Science Foundation of Shanghai under Grant No. 10ZR1421100, Innovation Program of Shanghai Education Commission under Grant No. 08YZ98, Postgraduate Construction of the Core Curriculum of University of Shanghai for Science and Technology under Grant No. 16000206 and Innovation Program of School of Optical-Electrical and Computer Engineering, USST, under Grant No. GDCX-T-102.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 303–311, 2010.

pruning rule. Fig.1 (b) depicts the TPL pruning method. $\perp(q, o)$ is the perpendicular of segment *qo*, according to the character of perpendicular, the objects in the shade are more closed to point *o* than to point *q,* so they will be pruned. This rule further improved the efficiency of RNN query. In recent years, researchers have done a lot of research work on R*k*NN query based on six-regions pruning and TPL pruning, we will give some introduction of the work in Section 2.



(a) Six-region pruning (b) TPL pruning

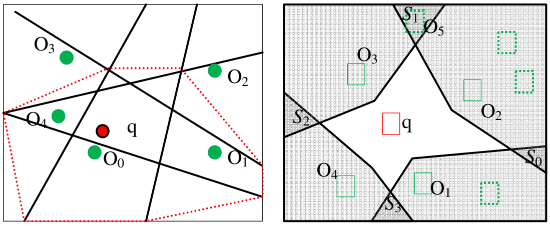
Fig. 1. Two popular pruning approaches

We find at least two shortages of existing algorithms of moving objects R*k*NN query. One is they only choose one *pruning reference object* and without the relationships between the *pruning reference object*s. This makes the pruning regions are too small to improve pruning efficiency. The other is some algorithms assign objects *safe regions* but they do not adaptively adjust the *safe regions*. If *safe regions* are too small, the mobile devices will report their location to servers and request re-computation frequently, which increases the cost of computation and communication. If *safe regions* are too large, the pruning regions may be small and the candidate set may be too large, that also leads to the increment of computation cost and the reduction of query efficiency. To overcome the drawbacks, we propose an **E**fficient and **S**table **A**pproach named ESA which includes two core algorithms: *k*-Binding algorithm and ASR algorithm. *k*-Binding is a pruning algorithm which selects *k* objects to be bound together as *pruning reference objects* for each time of pruning. ASR algorithm can adaptively adjust the *safe regions* for each *pruning reference object* and keeps the *safe regions* to be more suitable. In this way, ESA not only improves pruning efficiency but also reduces the communication cost and makes the server more stable.

2 Related Work

R*k*NN query is always used for searching k points which are more affected by one point. Wang et al [6] proposed R*k*NN evaluation algorithm called FINCH. Given an R*k*NN query and a set of data objects *S*, FINCH computes a region such that only the objects in the region and the objects in *S* can be the result objects .As shown in Fig.2. (a), the region of dashed polygon is the search region. The vertices of the polygon must meet the following two conditions: first, they must be the points in the intersections of TPL rule; second, all the regions contain them must be *k*+1 TPL lines pruned, or they are on the side of the search area. This algorithm is not suitable for moving objects very well, because its communication is expensive, and also its pruning cost is high for its large search region.

Cheema [7] presented an algorithm for continuous R*k*NN queries. In this algorithm, they assigned each object and query point with a rectangular safe region such that the expensive re-computation is not required as long as the query point and objects remain in their respective safe regions. This approach is called as *lazy updates* which significantly diminishes the computation cost. As a by-product, it also reduces the communication cost in client-server architectures because an object need not report its location to the server unless it leaves its *safe region* or the server sends a location update request. Fig.2. (b) shows us the *lazy updates* algorithm in R*k*NN query. When *k*=2, the pruning regions are S_0, S_1, S_2 and S_3 ,and only one object point O_5 is pruned at this time. If we want to get larger pruning regions, we should choose more *pruning reference objects*. This means the computation cost will be increased. On the other hand, it does not assign *safe regions* adaptively, which causes some regions so small that the objects will request updates frequently and some regions are so large that may wait a long time to updates. This brings the instability to the server.



(a) FINCH (b) Lazy updates

Fig. 2. FINCH and lazy updates methods

3 Our Approach

3.1 *k*-Binding Algorithm

To overcome the inefficiency of pruning for R*k*NN query in the filtering phase, *k*-Binding algorithm is proposed in ESA. To describe the algorithm more clearly, we do not assign *object points* with *safe regions* at first. The algorithm will be extended with the assignment of *object points* in Section 3.3. Then definitions and notations exploited in the following sections are shown in Table 1.

The main idea underlying *k*-Binding algorithm is presented as follows. First, find out the nearest neighbor of the *query point*, and then seek *k*-1 nearest neighbors which do not contain the *query point*. Second, bind these *k object points* as pruning reference points for TPL pruning respectively. Third, compute the intersection among the TPL pruning regions, these intersections are the pruning regions of this time of pruning. Fourth, repeat the steps above in the regions those not be pruned until the remaining *object points* are less than *k*. Finally, put the remainders into the candidate set.

Table 1. Definitions and notations used in this paper

Definitions	Explanation	Notations	Representation
object points	The *points,* including query requests and the points in the query rang	$NN(q, S)$	Denotes query R, the set of nearest points to *query point* q, in the object set S. $R=\{r\in S\mid dist(r, q) = mindist(S, q)\}$.
safe regions	regions moving objects keep moving in and do not affect RNN query results	$kNN(q, k, S)$	Denotes query R, the set of k nearest points to *query point* q, from *object points* set S. $R=\{r\in S\mid dist(r, q)\le dist(q, o)\}(\lvert R\rvert\le k)$.
query objects	*object points* in the query rang and returned as a member of query results	$RNN(q, S)$	Denotes query R, the set of reverse nearest neighbor of *query object* q in object set S. $R=\{r\in S \mid dist(r, q) = distNN(r, S)\}$.
pruning reference objects	*object points* selected as pruning references in filtering phase	$RkNN(q, k, S)$	Denotes query R, the set of k reverse nearest neighbors of *query object* q in object set S. R satisfies $R=\{r\in S \mid dist(r, q)< distkNN(r, S)\}$.
query points	*object points* stands query request		

Algorithm 1: k-Binding (q, k, S)

Input: *query point* q, value k, *query object* set S, initializing candidate set R: $=\varnothing$
Output: candidate set R
Description :

```
if $|S|\le k$ then
    $R \leftarrow R \cup S$;
    return $R$;
else
    $n \leftarrow NN(q, S)$;
    $R \leftarrow n$;
    $N \leftarrow k\text{-}1\ NN(n, k-1, S-q)$;
    For each object $n[i]$ in set $N$ do
        $H[i] \leftarrow TPL(q, n[i])$;
        $H \leftarrow H \cap H[i]$;
        $R \leftarrow n[i]$;
    $S \leftarrow S-R-H$;
    $k$-Binding ($q, k, S$);
```

Fig.3 shows how to implement R2NN query using k-Binding algorithm. Step 1, seek the nearest neighbor n_0 of the *query point* q, and then search the nearest neighbor (*1* nearest neighbor) n'_0 of n_0. Step 2, bind n_0 and n'_0 as pruning reference points to do TPL pruning with *query point* q. Step 3, compute the intersection of pruned regions. Step 4, find out the nearest neighbor n_1 in the regions not be pruned, and then query the nearest neighbor n'_1 of n_1, and together n_1 to do TPL pruning with *query point* q, and then compute the intersection. Step 4, find out n_2 and n'_2 to do TPL pruning with q and compute the intersection. Step 5, after above steps, there is only one object point n_3 remained, it is less than 2, so take n_3 into the candidate set R and then stop iteration. Finally, return the candidate set $R=\{n_0, n'_0, n_1, n'_1, n_2, n'_2, n_3\}$. In the figure, the shade area is the pruned region.

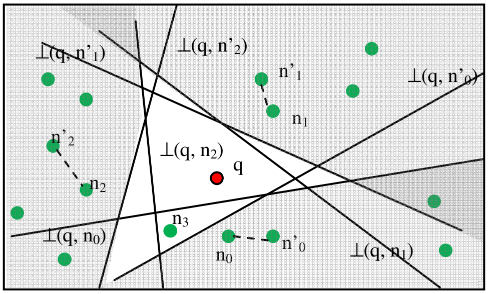


Fig. 3. *k*-Binding pruning

3.2 ASR Algorithm

ASR algorithm is another core algorithm of ESA and is used to assign *safe regions* suitably before query. By the employment of suitable *safe regions*, ASR reduces the communication cost and enhances the stability of the server.

In moving object R*k*NN query, as we know, if one of the *object points* moves out of its safe region, all the *object points* must report their locations to the server, and need to re-compute the R*k*NN. It reveals that communication cost and the frequency of re-computation depend on the moving object point that first moves out of its *safe region*. Therefore, we should take all *object points* as whole when we assign *object points* with *safe regions*.

Now, let us describe the main idea of ASR algorithm. The size of *safe region* is related to the maximum offset of the moving object. According to the maximum offset of *n* times of re-computation before and the time cost, we can easily predict the next average offset speed. Addition to the server acceptable re-computation time, we can obtain the size of rectangle sides of the next re-computation.

Here, we assume that *T* be the reasonable time to request a re-computation. *T* is called as *re-computation cycle*. v_0 denotes the initial velocity of *object point*. $S_{max}[i]$ denotes the maximum offset in *i*th re-computation cycle of moving object before, and $T_{cost}[i]$ denotes the time needed for moving maximum offset. *n* denotes latest *n* times re-computation cycle. The value of *n* is determined by the *accessing cycle C* and *re-computation cycle T*, $n \leq [C/T]$. The *safe region* side size of the next re-computation *L* satisfies Formula (5):

$$L = \begin{cases} 2Tv_0 (n = 1) \\ 4T\sum_{i}^{n}[\frac{S_{\max}[i]}{T_{\cos t}[i]} * \frac{n-i}{n(n+1)}](2 \leq n \leq \frac{C}{T}) \end{cases} \tag{5}$$

Where, (*n-i*)/*n* (*n+1*) is the weight of the maximum offset of *i*th re-computation cycle (counting of *n* begins with the latest cycle). According to the fraction, we get the re-computation cycle occurred near the larger weight it will be assigned. This is in the line with the facts. We describe ASR algorithm in Algorithm 2.

Algorithm 2: ASR (v_0, C, T, n)

Input: the initial velocity of *object point* v_0, the *accessing cycle* C, the re-computation cycle T, the re-computation time n
Output: the size of rectangle side L
Description:

```
1: $V_s \leftarrow 0$;   //initialize the maximum offset speed
2: if $n=1$ then $L \leftarrow 2TV_0$;
3: else
4:    if $n \geq [C/T]$ then $n \leftarrow [C/T]$;
5:    for each re-computation cycle i do
6:       $V_s \leftarrow V_s+(S_{max}[i]/T_{cost}[i])*((n-i) /n (n+1))$;
7:    $L \leftarrow 4T* Vs$;
8: return $L$;
```

One of the most important variables in Algorithm 2 is the maximum offset which we compute it using MaxOffset algorithm (See Algorithm 3).

The main task of MaxOffset algorithm is as follows: record the location and the time of the moving object when it changes its direction or runs out of its *safe region*; compute the offset value if the offset is bigger than before; change the maximum offset to this value; compute and record the consuming time and it. The description of MaxOffset algorithm is as below.

Algorithm 3: MaxOffset (l_0, t_0)

Input: the initial location l_0 and time t_0 of moving object in the *re-computation cycle*
Output: Maximum offset of re-computation cycle S_{max}, and its time consuming T_{cost}
Description:

```
1: $S_{max} \leftarrow 0$;
2: if the moving object changes its direction or any object runs out of its safe region
   then  //mobile device continuous detection
3:     $l_t \leftarrow$ the location at the moment
4:     $t_t \leftarrow$ the time at the moment;
5:     $S \leftarrow offset\ (l_t, l_0)$;
6:     $S_{max} \leftarrow max\ (S_{max}, S)$;
7:     if $S_{max}$ changed then $T_{cost}$: $\leftarrow t_t-t_0$;
8: if no object run out of safe region then
9:     repeat to step2;
10: return $S_{max}$ and $T_{cost}$;
```

3.3 Extending *k*-Binding Algorithm

In Section 3.1, we introduce *k*-Binding algorithm without assigning *safe regions* to *object points*. In this part, we extend the *k*-Binding algorithm by assigning *safe regions* to all the *object points*.

In the extended *k*-Binding algorithm, we choose rectangles as the shape of *safe regions*. Different for *k*-Binding, extended k-Binding algorithm moves the TPL lines to *pruning points* after doing the TPL pruning for each couple of *antipodal corners* in *pruning reference object* rectangle and *query object* rectangle. The *pruning point O* of antipodal corners *B* and *H* satisfies Equations (6) [7].

$$\begin{cases} o[i] = (R_{1L}[i] + R_{2L}[i]) / 2 (B[i] > H[i]) \\ o[j] = (R_{1H}[i] + R_{2H}[j]) / 2 (B[j] \leq H[j]) \end{cases} \quad (6)$$

In which, $\mathrm{R}_{1L}[i]$ and $R_{1H}[i]$ respectively denote the minimum and maximum coordinate values of *i*th dimension of rectangle R_1.

Fig.4 shows us an example of the extended *k*-Binding algorithm where *k*=2. In the figure, all the pruning points of antipodal corners overlap on a same point *O*, because of no overlapping dimensions between R_1 and R_2. All the lines in the figure intersect at the same point *O*; the lines move TPL lines of antipodal corners. The shade area of concentrated color is the pruning region. Although some antipodal corners' pruning points are different, the pruning region is also the intersection of all moved TPL pruning area. We needn't to give more detail here.

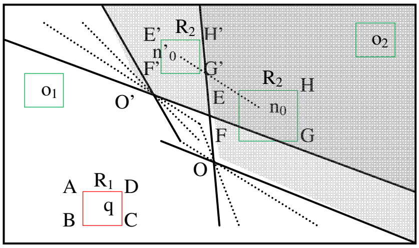


Fig. 4. Example of extended k-Binding

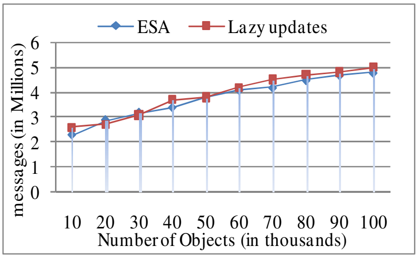


Fig. 5. Comparison of traffic load

3.4 Data Structure

In our system, there are two tables: *query object*s table and *query points* table. *Query objects* table stores the id and the safe region for each object. The *query points* table records the set of candidate objects for each query point.

We also use the Grid structure like the grid-based data structure in [7], Grid structure is better when updates are intensive because complex data structures such as R-tree and Quad-tree are expensive to update. So we choose Grid structure to store the locations and *safe regions* for moving objects and *query points*.

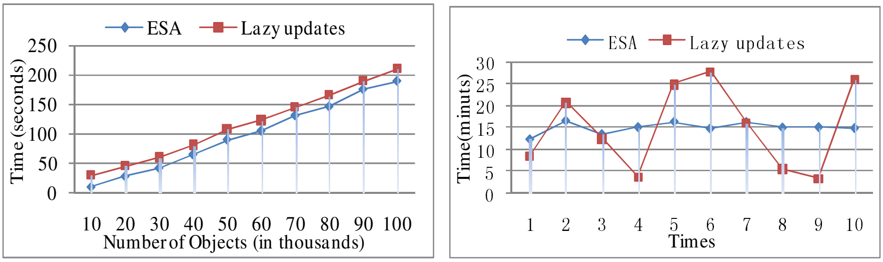
In our Grid date structure, each cell contains *object list* and *influence list*. Object list stores object ids of all objects whose *safe regions* overlaps the cell. This list is used to identify the objects that may be located in this cell. Influence list stores all *query points* ids for which this cell lies in the regions which are not pruned. The purpose of this is if an object moves into the cell, we can know that the queries in the influence list of this cell are affected. More detail about the grid structure is enumerated in [7].

4 Experiments

All the experiments were conducted on Intel Pentium 2.16 GHz dual CPU with 2 GB RAM. In existing methods for R*k*NN query, the most influential one is *lazy updates* proposed in [7]. Therefore, in our experiments, we compare ESA with lazy updates in communication cost, query efficiency and stability of server.

Fig.5. shows the comparison of the traffic load between *ESA* based R*k*NN query and *lazy updates based* R*k*NN query. The points of coordinates are the number of average messages in different number of *query objects*. The figure illustrates that the numbers of communication messages in *ESA* and *lazy updates* are approximate, which means they have similar effect in communication.

Fig.6 (a) depicts the computation cost comparison of R*k*NN query between ESA and lazy updates. The points of coordinates are the average time cost of R*k*NN query with different number of *query objects*. From the figure, we can draw that the time cost by ESA is obviously less than that by *lazy updates*, which indicates the ESA query is more efficient than *lazy updates* query.



(a) Comparison of Computation cost (b) Time interval between two query requests

Fig. 6. Comparison of Computation cost and Time interval

Fig.6. (b) describes the comparison of the time interval between two queries in 100,000 *query objects*. The *accessing cycle* is set to 150 (*C=150*) minutes and the re-computation cycle is set to *15* (*T=15*) minutes. The figure shows that the fluctuation of time interval of *lazy updates* is greater than that of ESA. This indicates ESA makes server more stable than lazy updates does.

From above experimental results we get that ESA is more efficient in R*k*NN query than *lazy updates*. The reason is ESA takes *k*-Binding algorithm to prune in filtering phase and takes ASR algorithm to adjust the *safe regions*, which makes the *safe regions* more suitable and the server more stable.

5 Conclusions

In this work, we propose an efficient and stable R*k*NN query approach, which we called it as ESA. ESA involves *k*-Binding algorithm and ASR algorithm. *K*-Binding algorithm improves the efficiency of R*k*NN queries by binding *k object points* as *pruning reference objects*. *ASR* algorithm makes the server more stable by adaptively adjusting *safe regions*. Our experiments show that ESA is better than *lazy updates* in improving query efficiency and enhancing the stability of server. In our future work, we will extend our approach to dichromatic query of R*k*NN which is also important in location based services.

References

1. Korn, F., Muthukrishnan, S.: Influence sets based on reverse nearest neighbor queries. In: The Proceedings of SIGMOD, pp. 201–212 (2000)
2. Stanoi, I., Agrawal, D., Abbadi, A.E.: Reverse nearest neighbor queries for dynamic databases. In: The Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2000)
3. Xia, T., Zhang, D.: Continuous reverse nearest neighbor monitoring. In: ICDE, p. 77 (2006)
4. Wu, W., Yang, F., Chan, C.Y., Tan, K.L.: Continuous reverse k-nearest-neighbor monitoring. In: Proceeding of MDM (2008)
5. Tao, Y., Papadias, D., Lian, X.: Reverse kNN search in arbitrary dimensionality. In: Proceedings of VLDB, pp. 744–755 (2004)
6. Wu, W., Yang, F., Chan, C.Y., Tan, K.L.: FINCH: Evaluating Reverse k-Nearest-Neighbor Queries on Location Data. In: The Proceedings of VLDB (2008)
7. Muhammad, A.C., Lin, X., Zhang, Y., et al.: Lazy Updates: An Efficient Approach to Continuously Monitoring Reverse kNN. In: Proceedings of VLDB (2009)

The Research of Efficient Dual-Port SRAM Data Exchange without Waiting with FIFO-Based Cache

Alfred Ji Qianqian, Zhao Ping, Cheng Sen, Tan Jingjing, Wei Xu, and Wei Yong

School of Electronics and Information, Northwestern Polytechnical University
710072, Xi'an, Shaanxi, P.R. China

Abstract. This paper proposes a program of efficient dual-port SRAM data exchange without waiting with FIFO-based cache, which is targeted for timely, massive and interactive features of data transmission in MIMO systems, using FIFO as a dual-port SRAM external cache to achieve real-time data exchange between multiple systems or processors. The program can solve time conflict and data covering problem in the competitive state of data storage, reduce the transmission delay to wait for data exchange. This paper uses dual-port SRAM CY7C019 to do a simulation test for the program, which can realize effective addressing between memory and CPU in address mapping way. By the analyses to the system performance, the effectiveness and feasibility of this program is proved.

Keywords: dual-port SRAM, FIFO buffer, data exchange, address mapping.

1 Introduction

Large data collection and high-speed transmission is an important part of communication system, multiple-input multiple-output (MIMO) which has high speed characteristics of low-cost data exchange becomes one of the key technologies of new communications era. As the data exchange transmission capacity is enormous in MIMO system, multi-port memory can be used to simplify the data exchange structural complexity of data receiving and sending module, which can increase data transfer rate, bring about reading and writing simultaneously in parallel exchange for multiple processors, and improve data transfer rate on MIMO systems to some extent[1]. Because there are some of the flaws of multi-port memory, the efficient dual-port SRAM program based on first in first out(FIFO) being for buffer is proposed in this paper, which can avoid the time conflicts between of reading and writing when multi-ports have data exchange at the same time, and also solve data exchange delay, reading and writing coverage problems, so as to improve the use of the dual-port memory in the MIMO system and achieve an efficient writing operation without waiting.

2 Conventional Dual-Port SRAM Problems

Dual-port SRAM can read and write internet data in two-way without external control logic that is of better data exchange performance. When there is data exchange

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 312–319, 2010.

between double systems interoperability, memory can be used as the common shared memory for data exchange to provide transmission channel for the complex data exchange. Dual-port SRAM allows two ports visit the same storage unit at the same time, but if the read and write operations are both on the same storage unit at the same time, the competition of arbitration between reading and writing will be lead, the waiting for the data access operations embedded by clock cycles will be result in, and the system data exchange process will be affected. If one port is writing, meanwhile, the other port is reading, the readout in the read cycle may be old data, it also may be new data. Write cycles always use new data to update storage units, so when two ports have write operation to the same memory cell at the same time, it will certainly lead to a competition problem of metastable state of memory cells.[2]

Because of the competition, data transmission of two ports in the Dual-port SRAM is easy to generate data reading and writing conflict, impact on data transfer rate, result in data processing “bottlenecks”, so it cannot meet the design requirements of real-time speed data exchange.

3 An Program of Efficient Dual-Port SRAM Data Exchange without Waiting with FIFO-Based Cache

From the simulation and comparison on reading power and operating frequency to the 8T SRAM, 10T and 10T differential single-ended SRAM and so on by Hiroki Noguchi [3], we can see that: as the external circuit complexity increasing, the reading power that memory consumption will increase or operating cycle time will be longer, which proves there are some deficiencies of the SRAM in the traditional applications. An improved strategy of the SRAM being of improvement has been proposed in related literature [4], and the concept of using FIFO as buffer was referred to in the design process, but only as an internal memory cell in the SRAM, at last this program using the SRAM as the shared memory confirmed the advantages of that the original speed of the ATM was increased massively. However, to a large number of frequent data exchange in the MIMO system, the method is in the requirement for the additional cache higher capacity, and also its interior design is more complex, besides, it is necessary to insert waiting state for the competitive conflict of both sides. So this article proposes a thought to achieve the external cache of SRAM with FIFO, and then researches the program of efficient dual-port SRAM data exchange without waiting with FIFO-based cache to improve the ability of data exchange between the multi-systems or processors.

It is very easy to apply FIFO as the data cache because of FIFO’s characteristics, which can read and write without external address line[5]. We adopt FIFO to be cache to improve the data exchange rates of dual-port SRAM, the system circuit connection scheme is shown as figure 1.

The key components of the system circuit program are: dual-port SRAM CY7C019 (128*9bit), 6 FIFO memory CY7C429 (2k*9bit), 1 address latch 74ls373 and 2 CPU AT89c51. Among them, CPU A and CPU B are the host to transfer data, it controls the data transfer process and achieve the transient.

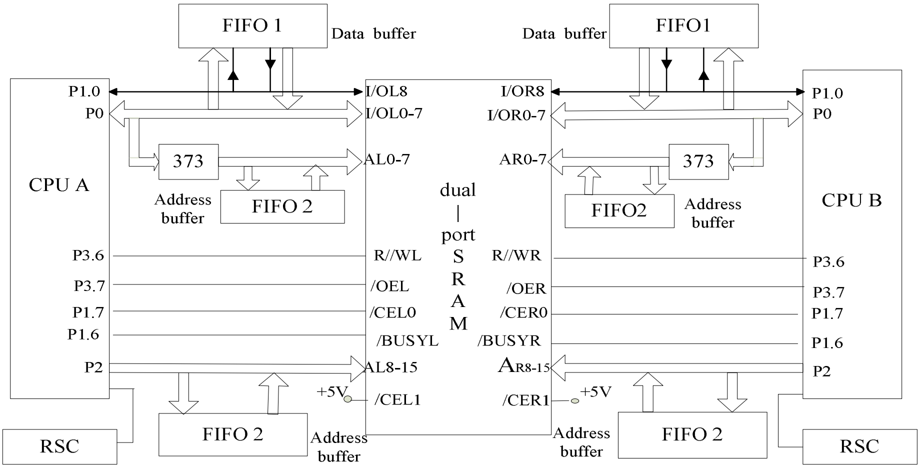


Fig. 1. The circuit diagram of dual-port SRAM with FIFO-based cache

4 Hardware Operations Timing

Dual-port SRAM is used as the shared memory to achieve both interoperability of data between the CPUs, the arbitration operation timing diagram of SRAM CY7C019 is shown as Figure 2. The timing diagram can be described as: address A0-A15 are maintained a minimum time of trc; CE chip selection is given to the LEFT port first to achieve the response from the visit of the LEFT port when two ports visit the same address unit; at the same time the RIGHT port visits the same address which the LEFT port visits, then the /BUSY is effective; when the access of the LEFT port is over, the /BUSY of the RIGHT port is invalid; as the /BUSY is invalid, RIGHT port begin to visit; because the visit of RIGHT port has already begun, the /BUSY of LEFT port is effective, then turn to the next round of data manipulation orderly.

FIFO as for the buffer of Dual-port SRAM is used to store the data being sent to prevent the data loss or transmission delay in competitive state, it can memory the address data of RAM which is operated in busy state ,in order to prepare for the data

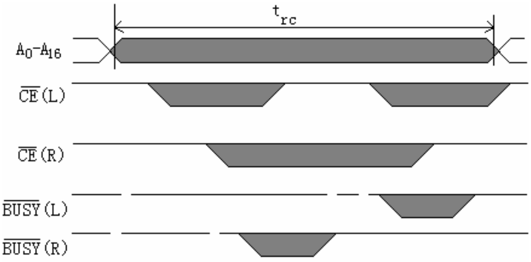


Fig. 2. The arbitration operation timing of CY7C019

output of RAM from the data register FIFO, the timing of CY7C429 data operation is shown in Figure 3, the operation sequence is described as: when /EF is '1 ', FIFO is not empty, /R is given a low level, Q0-Q8 start to read the data, once access time of reading is 10ns, the read operation is over as /EF converted high level into low level, and then FIFO is empty; /W is set to low level, then writing data is to be started, FIFO is not empty, once write time is 6ns; /W is converted from low level to high level after FIFO has been filled, and then the write operation is over, /FF is '0 '; /R is given low level, and next round of read operation is to bugun.

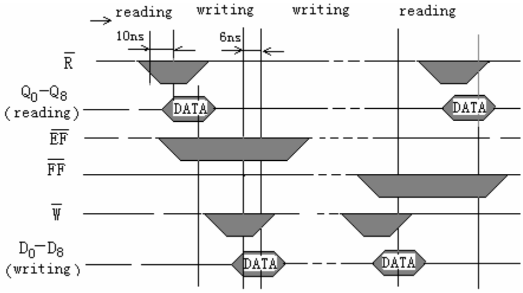


Fig. 3. The access and labeling operation of CY7C429

5 System Workflow

System flow chart is shown as figure 4.

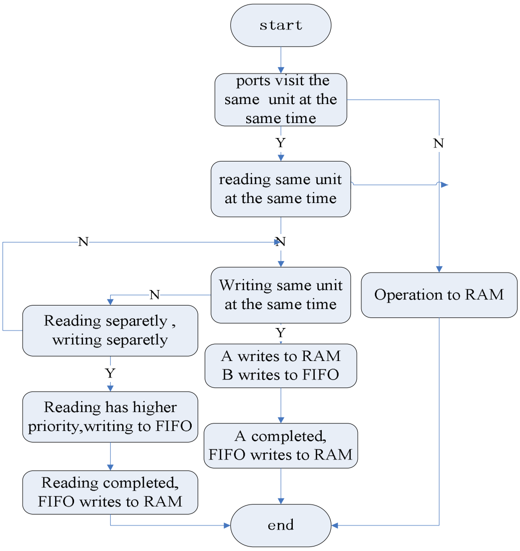

start
ports visit the same unit at the same time
Y
N
reading same unit at the same time
N
N
Writing same unit at the same time
N
Operation to RAM
Reading separetly , writing separetly
Y
Y
A writes to RAM B writes to FIFO
Reading has higher priority,writing to FIFO
A completed, FIFO writes to RAM
Reading completed, FIFO writes to RAM
end

Fig. 4. System workflow chart

Principle of System workflow as follows:

(1) When the time-sharing operation of single reading and single writing to the SRAM are happened by CPU A and CPU B separately, the output of pin 1.7 is low level '0', so the pin /CEL0(when CPU A is on operation) or the pin /CER0 (when CPU B is on operation)connected with pin 1.7 is active because of its low level, at now the SRAM will be selected, then the one-direction read/write operation between CPU and SRAM is accured. When one port is undertaking reading/writing, the /BUSY which is of the other port is set 1 to be idle state, and pin /R of FIFO buffer memory in the corresponding port is active, but the empty tag /EF of internal memory unit in FIFO is '0', so the read operation of FIFO is forbidden, and the FIFO buffer memory is in inoperation condition.

(2) When CPU A and CPU B implement the reading operation to the dual-port SRAM at the same time, the signal output line pin 1.7 of both ports is set be low, the pin /CEL0(when CPU A is on operation) or pin /CER0 (when CPU B is on operation) is active because of its low level, then system searches the appropriate data storage unit in the dual-port SRAM with the address stata of CPUs and reads data of each storage unit according to data pointers decreasing, the /BUSY signal lines of SRAM are both set be '0', the pin /R of FIFO buffer memory in the corresponding port is active, but the empty tag /EF of internal memory unit in FIFO is '0', and the FIFO buffer memory is still in inoperation condition.

(3) When CPU A and CPU B implement the writing operation to the dual-port SRAM at the same time, both of them send writing request signal to dual-port SRAM at first, only one CPU A(or CPU B) is answered by the analysis and judgment of dual-port SRAM, and the R//W is set ‘1’ to be active, then the data is written into the dual SRAM internal memory unit. At the same time the /BUSY of the other port of SRAM is set '0', and /CS is set '1' to be inactive, then the pink /W of FIFO linked to /BUSY of SRAM is set '0' to be active, the data which will be transferred to the SRAM from this CPU is written into data FIFO buffer memory and the addressing data which the operation happened to in the SRAM is written into addressing FIFO buffer memory, the /EF is set '1', which shows the FIFO is not null, and then the writing operation of FIFO buffer is completed. After the operation between CPU and dual SRAM is accomplished, the /BUSY is set '1' to be idle, then the pin /W of FIFO is set high level to be inactive, and /R is set low level to be active. The /CS of this port of SRAM is set '0' to select the SRAM and write the data held in FIFO to the matching address unit of dual SRAM according to the addressing data held in FIFO to complete the whole reading process.

(4) One CPU reads data from dual SRAM, while the other writes data into dual-port SRAM, the read operation has higher priority. When the write-state CPU receive the busy signal of the other CPU, the chip select signal /CS is set '1' to be inactive, which means the write-state CPU don't work with dual SRAM but write the data into FIFO buffer temporarily, the procedures of the writing resembles the buffer read/write procedure in (3).

Box structure [6] because of its sub-box and sub-grid form, many storage unit is needed; the allocation and management to the right to use the shared memory area by communication pool structure [7] is more complex, and the communication speed is slower. All of which cannot meet the requirement for the allocation and management to the right to use the shared memory area with this program, so this paper adopts the centralized control mode (distribution and management are focused on a unit), allocates and manages memory area which is of dual-port SRAM is main memory, and FIFO is assistant memory by combining the soft and hard to focus main storage area and assistant storage area into an unified whole. All access space of dual-port memory is 128k*9 bit, and memory space of FIFO buffer is 2k*9 bit, the data bus and address bus of dual-port SRAM and FIFO buffer can be linked to those of CPUs together, chip select /CE0 is linked with page strobe line /PAGEO, which means the start mapping address is 0000H[8], and the mapping space were 00000H-1FFFFH, 20000H-208FFH separately.

6 Performance Simulation and Comparative Analysis

Take the competitive state that two ports write the same address unit simultaneously for example, the competitive timing diagrams of conventional system and improved system are shown as Figure 5 and Figure 6 respectively:

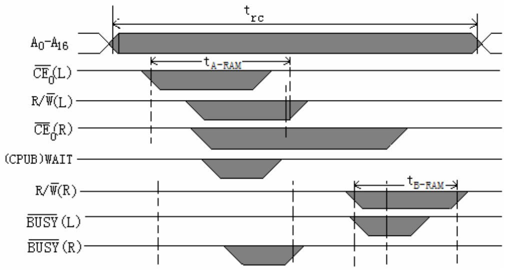


Fig. 5. The competed timing diagram in the conventional dual-port SRAM system application

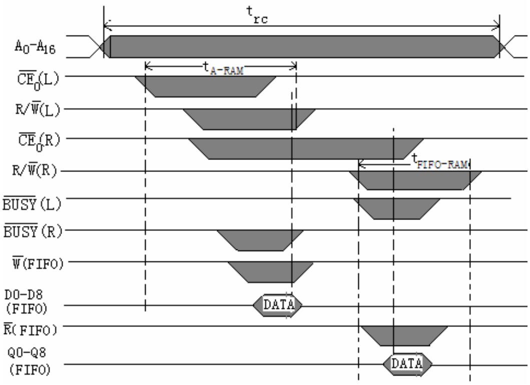


Fig. 6. The competed timing diagram of improved system

Assume that there is 2K data writing to SRAM, the conventional dual-port SRAM system with the improved dual-port SRAM system is compared:

The transmission can be divided into the operation of A and the operation of B separately in the conventional system, whose whole time $t^{brc} = t^{A\rightarrow RAM} + t^{B\rightarrow RAM}$;

After improvement, the reading of CPU A is responded and the writing which is of CPU B to FIFO is happened at the same time, then data is written to SRAM from FIFO when the operation of CPU A is completed, and the reading time of FIFO is faster than that of SRAM, so $t^{B\rightarrow FIFO}$ will be covered by $t^{A\rightarrow RAM}$, and the whole time of transmission $t^{arc} = t^{A\rightarrow RAM} + t^{FIFO\rightarrow RAM}$.

Here the oscillation frequency of 12MHZ of AT89C51 is adopted, and the instruction cycle when data is sent to external memory is 24 oscillating periods, so once reading time is $t^{once} = 24\times\frac{1}{12\text{ MHZ}} = 2\times10^{3}$ ns,

$t^{A\rightarrow RAM} = t^{B\rightarrow RAM} = 2\text{KB}\times t^{once} = 32768000\text{ns}$,

$t^{brc} = t^{A\rightarrow RAM} + t^{B\rightarrow RAM} = 65536000\text{ns}$,

The once reading time of the FIFO chosen in this paper is $t^{once} = 10$ns.

$t^{\text{FIFO}\rightarrow RAM} = 2\text{KB}\times t^{once} = 20480\text{ns}$,

$t^{arc} = t^{A\rightarrow RAM} + t^{FIFO\rightarrow RAM} = 32788480\text{ns}$,

Compared the conventional system with the improved, we can get the ratio of the time saved is $\eta = \frac{t_{brc} - t_{arc}}{t_{brc}} \approx 50\%$.

According the timing calculation, we can see the advantages of the improved dual-port SRAM system reflected in the data transmission speed is greatly increased by 50% more or less, the delay waiting is reduced to meet the features of real time without waiting for the MIMO system.

7 Conclusion

In this system, FIFO storage area is set as associate storage area of the system, which increases the capacity of storage area of the system, provides a buffer for the data manipulation of the main storage area, improves the work process of the dual-port SRAM, and increases the data exchange rate effectively to implement Real-time system data exchange of MIMO system. The use of FIFO buffer not only increases reliability and simplifies the complexity of system structure, but reduces the time of system design and debugging, thus the cost of the system is saved.

References

[1] Li, J., Conan, J., Pierre, S.: Mobile Station Location Estimation for MIMO Communication Systems. In: 3rd International Symposium on Wireless Communication Systems 2006, ISWCS 2006, pp. 561–563 (2006)

[2] Liu, L., Nagaraj, P., Upadhyaya, S., Sridhar, R.: Defect Analysis and Defect Tolerant Design of Multi-port SRAMs. Journal of Electronic Testing: Theory and Applications (JETTA) 24(1-3), 165–179 (2008)

[3] Noguchi, H., Okumura, S., Iguchi, Y., Fujiwara, H., Morita, Y., Nii, K., Kawaguchi, H., Yoshimoto, M.: Which is the Best Dual-port SRAM in 45-nm Process Technology? -8T, 10T Single End, and 10T Differential. In: Proceedings - 2008 IEEE International Conference on Integrated Circuit Design and Technology, ICICDT, pp. 55–58 (2008)

[4] Yang, H.-I., Chang, M.-H., Lin, T.-J., Ou, S.-H., Deng, S.-S., Liu, C.-W., Hwang, W.: A Controllable Low-Power Dual-Port Embedded SRAM for DSP Processor. In: Records of the IEEE International Workshop on Memory Technology, Design and Testing, pp. 27–30 (2007)

[5] Li, F.: Fairness Analysis in Competitive FIFO Buffer Management. In: Conference Proceedings of the IEEE International Performance, Computing, and Communications Conference, pp. 239–246 (2008)

[6] Jiannong, C., Xinyu, F., Lu, J., SajalK, D.: Mailbox-based scheme for mobile agent communications. Computer 35(9), 54–60 (2002)

[7] Xiaofei, X., Yi, L., JiChang, K.: The Research on Zero-Copy Receiving Method Based on Communication-Page Pool. In: Proceedings Parallel and Distributed Computing, Applications and Technologies, PDCAT, pp. 416–419 (2003)

[8] Jing, W., Xiaoya, F., Hai, W., Ming, Y.: A 16-Port Data Cache for Chip Multi-Processor Architecture. In: 2007 8th International Conference on Electronic Measurement and Instruments, ICEMI, pp. 3183–3186 (2007)

Extracting Service Aspects from Web Reviews

Jinmei Hao[1], Suke Li[2], and Zhong Chen[2]

[1] Beijing Union University, China
haomei99@yahoo.com.cn
[2] School of Electronics Engineering and Computer Science, Peking University, China
{lisuke,chen}@infosec.pku.edu.cn

Abstract. Web users have published huge amounts of opinions about services in blogs, Web forums and other review friendly social websites. Consumers form their judgements to service quality according to a variety of service aspects which may be mentioned in different Web reviews. The research challenge is how to extract service aspects from service related Web reviews for conducting automatic service quality evaluation. To address this problem, this paper proposes four different methods to extract service aspects. Two methods are unsupervised methods and the other two methods are supervised methods. In the first method, we use FP-tree to find frequent aspects. The second method is graph-based method. We employ state-of-the-art machine learning methods such as CRFs (Conditional Random Fields) and MLN (Markov Logic Network) to extract service aspects. Experimental results show graph-based method outperforms FP-tree method. We also find that MLN performs well compared to other three methods.

Keywords: service aspect extraction; opinion mining; web mining.

1 Introduction

Consumers are used to browsing online service related reviews and like to compare service quality of different service providers before making purchasing decisions in recent years. It is common for Web uses to publish Web reviews to describe a variety of service aspects of services due to convenient facilities of Web based service providers such as *tripadvisor*[1].

Service quality evaluation plays a very important role in traditional service management. Service quality has five dimensions [1] : reliability, responsiveness, assurance, empathy and tangibles. In this paper, we try to extract five dimensions related service aspects. Each dimension has different aspects. For example, in tangible dimension includes aspects: the appearance of physical facilities, equipment, and also other tangible evidence of the care and attention to detail that are exhibited by service providers. For instance, in the sentence "*The bathroom is great.*", *"bathroom"* is a service aspect.

The research challenge is how to extract service aspects from service reviews in order to conduct automatic service quality evaluation. To address this issue,

[1] http://www.tripadvisor.com

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 320–327, 2010.

this paper proposes four different methods to extract service aspects. The first two methods are unsupervised methods and the last two methods are supervised methods. In the first method, we use FP-tree [2] to find frequent aspects. The second method is graph-based method. We employ CRFs (Conditional Random Fields) [3] and MLN (Markov Logic Network) [4] as the third and the fourth methods. Experimental results show graph-based method outperforms FP-tree [2] based method. We also found MLN performed well compared to other three methods.

There are several potential applications based on extracted service aspects. For instance, we can find what service aspects may affect purchasing behaviors of customers. And what service aspects make customers feel good or bad. We can also summarize reviews uses sentences containing the most important aspects. It is possible to rank the helpfulness of the reviews according to the extracted aspects. In service management research field, researchers usually design questionnaires or surveys to find the gap between customer expectations and perceptions. However, how to design proper questionnaires or surveys becomes a big challenge, because the designer needs to know what service aspects are important for most of customers. Efficient service aspect extraction methods are promising for survey design.

2 Related Work

Few publications focus on service aspect extraction. However, there are extensive research work focusing on product feature extraction. Hu and Liu's work [5] is early effort to summarize product opinions through association mining. Liu [6] proposed an opinion mining system to do feature-based opinion mining. In [5] and [6], nouns and noun phrases were used as product features. There are also some research focusing on implicit feature extraction, such as [7]. However, our work focuses on explicit service aspects only. Popescu et al. [8] introduced an unsupervised information extraction system, namely OPINE which parses the reviews and applies a simple pronoun-resolution module to the parsed data. It is not clear whether OPINE can been applied to service aspect extraction or other languages. Since our work only adopts shallow language processing techniques, the proposed methods are quite general and can be applied to other languages easily.

3 Service Aspect Extraction

Service aspect extraction is the first step toward further service opinion mining in our work. In this section, we show four methods to extract service aspects respectively. The first two methods are unsupervised methods, and the last two methods are supervised methods using CRFs and MLN to extract service aspects. In this work, service aspects are nouns and noun unit (consecutive nouns) that are related service quality evaluation. For example *hotel*, *room*, *breakfast*, *staff*, *location*, etc. are all service aspects. We judge whether nouns or noun units are service aspects by humans.

3.1 Association Mining Based Aspect Extraction

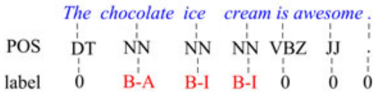
Hu and Liu's work [5] is the early effort for feature-based product opinion mining using association mining algorithm Apriori [9]. Hu and Liu [5] believed when people comment on product features, the words that they use converge. Because most of service aspects are nouns or noun phrase, intuitively Hu and Liu's approach [5] can also be applied to our data set. In this paper, we take hotel industry as an example to extract service aspects. However, a sentence can have opinions about several service aspects and not every aspect is explicit. An aspect is explicit, if the service aspect words or phrases appear in the sentence. In this work, we only consider to extracting explicit service aspects.

3.2 Aspect Extraction Based on Conditional Random Fields

Conditional Random Fields (CRFs) [3] is used widely in sequence labeling. It is a framework for building probabilistic models to segment and label sequence data based on undirected graphical models. CRFs are also been applied in open information extraction [10] as well as POS tagging and phrase chunking. Suppose input data sequence is X and the label sequence is Y, then the joint distribution of Y given X is

$$p(y|x) \propto exp(\sum_{e \in E,k} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V,k} \mu_k g_k(e, y|_v, x)), \tag{1}$$

where, x is a data sequence, y a label sequence. Function f_k and g_k are feature functions which can acquire from training data, and $y|_S$ is the set of components of y associated with the vertices in subgraph S. To obtain labeled data for CRFs, in this method, the beginning noun of a noun phrase is labeled as ***B-A***, other sequential nouns are labeled as ***B-I***. A training example for CRFs is as Fig. 1 shown.



	The	chocolate	ice	cream	is	awesome	.
POS	DT	NN	NN	NN	VBZ	JJ	.
label	0	B-A	B-I	B-I	0	0	0

Fig. 1. A label sequence example for CRFs

3.3 Graph-Based Service Aspect Extraction

We use terms of sentences to construct directed graph. In order to simplify the process, we adopt a simple arc generation rule to generate graphs for service review. The rule is simple:

Arc Generation Rule. *If two terms in a sentence, for example term A and term B are consecutive, term A is term B's left neighbor, and if there is no arc from term A to term B, then an arc is generated from term A to term B.*

All service reviews in the data set will generate a directed graph, namely *term-tag* graph. Nodes in such a directed graph are terms which are extracted according to POS (Part-of-Speech) tags of sentences, and a tag is a attribute of a node. Each node is corresponding with one distinct term. It is obvious that even for the same term, it is possible that the term can have different POS tags due to the characteristics of natural languages. In this method, a node in a generated graph can have one tag attribute. If a term has more than one kind of tag, the tag with the most frequency will become the tag attribute of the node. For instance, the word *good* may have different POS tags in terms of different context. In the sentence *"He is a good man"*, the word good is labeled as ADJ. However, in the sentence *"I knew it was no good to say anything"*, the word *good* is a noun and should be labeled as NN. In our data set, because the frequency of tag ADJ is much bigger than the frequency of tag NN for the term *good*, the tag attribute of the term node of *good* is ADJ. In the opinion mining field, nouns and noun phrases are often used as candidate features or aspects. Intuitively, a major service aspect may be mentioned in different reviews and may have more in-links from other nodes or out-links to other nodes. Based on this observation, we use PageRank algorithm PageRank [11] on our constructed term-tag graph to rank all the terms. We then select noun or noun phrase terms as extracted service aspects according to their rankings. We consider the service aspect ranking problem as finding the top K most influential nodes in a term-tag graph. Let G be directed term-tag graph. The term ranking process is a random surfer process. Matrix A is the adjacency matrix or stochastic transition matrix of G which has n nodes. Let P is a n-dimensional column vector of aspect-opinion rank values:

$$P = (p_1, p_2, ..., p_n)^T \tag{2}$$

Matrix A is the adjacency matrix with

$$A_{ij} = \begin{cases} \frac{1}{O_i} & if(i,j) \in E, \\ 0 & otherwise \end{cases} \tag{3}$$

where O_i is the out-degree of node i. Node i can be a term. We can iterate to obtain rank values of all nodes them using

$$P = (1-d)e + dA^T P \tag{4}$$

where $d \in (0,1)$, d is a damping factor, and e is a column vector of all 1's. In our experiments, d is 0.85. For a random surfer in the graph, it has d probability to follow an out-link of the node and $(1-d)$ probability to jump to a random node. Fig. 2 is a term-tag example graph.

3.4 Aspect Extraction Base on MLN (Markov Logic Network)

Markov Logic Network (MLN) [4] combines first-order logic and probabilistic graphical models. A MLN can be viewed as a template for constructing Markov Random Fields,

$$P(X = x) = \frac{1}{Z} exp\langle \sum_i w_i n_i(x) \rangle = \frac{1}{Z} \prod_i \phi_i(x_i)^{n_i(x)}, \tag{5}$$

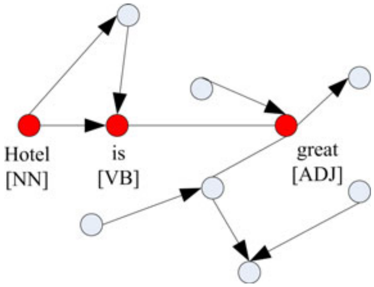


Fig. 2. A term-tag example

where $n_i(x)$ is the number of true groundings of F_i in X, x_i is the state (true values) of atoms appearing in F_i. In this method, a term is classified into two classes: aspect class and non-aspect class using MLN.

Table 1 contains all predicates that are used by MLN. For instance, $isSeqNVA(x, y)$ means term x and term y are in a subsequence of a tag sequence of a sentences. The tag sequence is $< NP, VP, ADJP >$. After we construct MLN, we can answer probabilistic queries such as $isAspect(x)$ which means the probability of x is a service aspect. In this work, if we get $p(isAspect(x)) > 0.5$, then we deem that x is a service aspect.

Table 1. Predicates and Their Descriptions

Rule	Description
isAdj(x)	X is an adjective.
isNon(x)	X is a noun.
isModifier(x,y)	X is a modifier of y.
isSeqNVA(x,y)	X and y are in the sequence of $< NP, VP, ADJP >$.
isInNPPhrase(x)	X is in NP phrase.

MLN needs to employ some logic formulas to work. These formulas are shown as follows.

$$\forall x(isNon(x) \wedge isAdj(y) \wedge isModifier(x, y)) \rightarrow isAspect(x)$$
$$\forall x(isNon(x) \wedge isAdj(y) \wedge isSeqNBA(x, y)) \rightarrow isApsect(x)$$
$$\forall x(isNon(x) \wedge \neg isInNPPhrase(x) \rightarrow \neg isAspect(x)$$
$$\forall x(isNon(x)) \rightarrow \neg isAdj(x)$$
$$\forall x(isAdj(x)) \rightarrow \neg isAspect(x)$$

4 Experiments

4.1 Experiments with Aspect Extraction

Our data set contains 500 reviews that are uniformly randomly sampled from the global data set without replacement. The global data set contain about 25

Table 2. Data Set Description

Method	Precision	Recall	F-Score
Pure Noun	0.4174	0.9024	0.5708
NP Noun	0.6172	0.4059	0.4898

thousands reviews crawled from *Triadvisor*[2]. The labeled data set contains 500 reviews and it has been segmented into 4545 sentences. The number of distinct labeled aspects is 1706 and they are distributed in 3421 sentences. These aspects are judged by humans. The work of labeling service aspect is labor intensive and time consuming. We also use OpenNLP[3] to get all the Part-of-Speech tags for the sentences. Noun and noun phrase can be used as features [5]. Table 2 shows some results of using *Pure Noun* and *NP Noun* as service aspects. *Pure Noun* means we only consider nouns only. *NP Noun* means we use nouns in noun phrases only. We can see if we use pure noun as service aspects, the precision is only 0.4174. NP Noun has higher precision but lower recall.

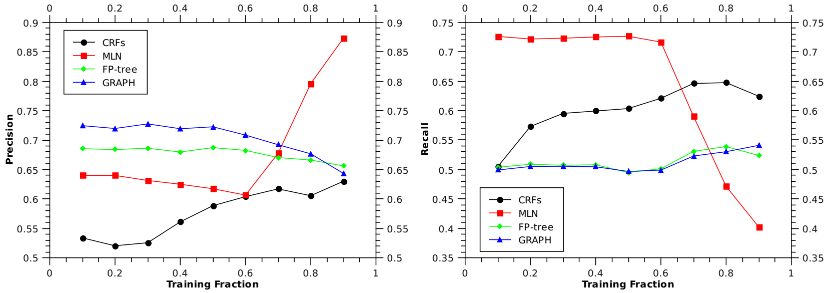
Because the graph-based method and FP-tree method are unsupervised methods, and the CRFs method and MLN method are supervised methods, in order to make comparison, graph-based method and FP-tree method only use test data set to extract service aspects in our experiments. The same as [5], this work uses support value 0.01 to extract product features. For instance, in Fig. 3a, when training fraction is 0.2, graph-based method only extracts aspects from the test data set with fraction of 0.8. In our experiments, we adopt 0.01 as our support value for FP-tree. For the graph-based method, we first use FP-tree on test data set with support 0.01 to get the number of extracted aspect, namely K, than we select K ranked aspects generated from graph-based method. However, CRFs method and MLN method are supervised machine learning methods and need training data set to work. We adopt CRF++[4] for labeling sequential data.

Fig. 3a shows the precision distributions of different methods. Graph-based method performs best. However, CRFs method performs worst. With the increase of training fraction, the precision also increases except for MLN method. Fig. 3b illustrates distributions of recall of four methods. We can see CRFs based method perform better than MLN method. MLN method has good performance when the training fraction less than 0.6. Fig. 4 shows distributions of f-score of our methods. In this case, FP-tree has the worst performance when training fraction greater than 0.4. F-score of CRFs based method increases with the increase of training fraction. When the training data fraction is less than 0.7, MLN method has the best performance.

[2] http://www.tripadvisor.com

[3] http://www.opennlp.org

[4] http://crfpp.sourceforge.net/



(a) Precision distributions (b) Recall distributions

Fig. 3. Precision and recall distributions

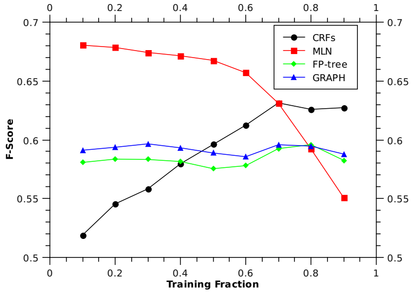


Fig. 4. F-Score distributions

5 Conclusions and Future Work

This work focuses on service aspect extraction. We propose four methods to conduct service aspect extraction. In the first method, we use FP-tree to find frequent aspects. The second method is graph-based method. We employ CRFs (Conditional Random Fields) and MLN (Markov Logic Network) as the third and the fourth methods. For measuring extraction precision, experimental results show graph-based method outperforms FP-tree based and other two methods in almost all cases. MLN method performs well in measuring extraction recall. In measuring F-score of service aspect extraction, We also find MLN outperforms other three methods when the fraction of training data set is less than 0.7. In the future, we will continue our research work toward automatic service quality evaluation based on extracted service aspects.

References

1. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: Servqual: A multiple-item scale for measuring consumer perceptions. Journal of Retailing 64 (1988)
2. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, SIGMOD 2000, pp. 1–12. ACM, New York (2000)

3. Lafferty, J.M.A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. the 18th International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2001)
4. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning 62, 107–133 (2006)
5. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 168–177 (2004)
6. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: Proceedings of the 14th international conference on World Wide Web, WWW 2005, pp. 342–351 (2005)
7. Su, Q., Xu, X., Guo, H., Guo, Z., Wu, X., Zhang, X., Swen, B., Su, Z.: Hidden sentiment association in chinese web opinion mining. In: Proceeding of the 17th international conference on World Wide Web, WWW 2008, pp. 959–968. ACM, New York (2008)
8. Popescu, A.M., Nguyen, B., Etzioni, O.: Opine: extracting product features and opinions from reviews. In: Proceedings of HLT/EMNLP on Interactive Demonstrations, pp. 32–33. ACL, Morristown (2005)
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB 1994, pp. 487–499 (1994)
10. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Commun. ACM 51(12), 68–74 (2008)
11. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30(1-7), 107–117 (1998)

Clustering the Tagged Resources Using STAC

Feihang Gao, Kening Gao, and Bin Zhang[1]

College of Information Science and Technology
Northeastern University
Shenyang, China
gaofeihang@163.com, gkn@cc.neu.edu.cn, zhangbin@ise.neu.edu.cn

Abstract. Similarity calculation is a key step in the process of clustering. Because most tagged resources on the Internet lack text information, traditional similarity measures cannot obtain good results. We propose the STAC measure to solve the problem of calculating the similarity between tagged resources. In the calculation of STAC, the similarity between tags is calculated using tag co-occurrence information, and the similarity between tagged resources is calculated based on tag comparison. Experiments show the clustering results of tagged resources using STAC is significantly better than using other traditional metrics such as the Euclidean distance and Jaccard coefficient.

Keywords: tagged resources; similarity measure; clustering.

1 Introduction

Tag is a typical feature of Web 2.0 and widely used in recent years. The contribution of collaborative tagging service and folksonomy [1] makes almost all the resources on the Internet, such as blog posts, videos, images, shared files and bookmarks carry category information in form of tags. But when the amount of these tagged resources becomes very large, it will be difficult for the user to find the wanted resource through searching tags. Clustering can help to solve this problem. It divides the resources into groups according to the similarity among them. Then the user can explore the resources more efficiently through resource groups.

Though the clustering algorithms available are almost perfect, but for the tagged resources on the Internet, the calculation of similarity becomes a big problem. These resources differ from traditional web documents. Most of them only have tags as the unique kind of attributes that can be used. If the similarity is not calculated correctly, the clustering result can not be assured, no matter how excellent the clustering algorithm is. Therefore, how to use the limited tag information to calculate the most accurate similarity is a problem worthy of study.

An important feature of tags is that a group of them describe the same object. If two tags appear in the description of the same object, they will be somehow relevant to each other. This kind of relation between tags is well-known as co-occurrence. The co-occurrence of tags is an expression of semantic relation. It can be regarded as a

[1] Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 328–335, 2010.

measure of similarity. The similarity between tags is very helpful to the calculation of tagged resources. Without this information, each resource can only be denoted by a sparse Boolean vector. The calculation of Euclidean distance and Jaccard coefficient based on these vectors will be inaccurate.

In this paper, we analyze the similarity between tags based on co-occurrence information and propose the STAC (Similarity based on TAg Comparison) measure which calculates the similarity between tagged resources through comparing tags and accumulating the similarity between tags. Excellent result is obtained in the clustering of tagged resources when STAC is used as the similarity measure.

This paper is organized as follows. In section 2 we introduce the related work. The STAC measure is proposed in section 3. Section 4 shows how STAC outperforms traditional similarity metrics through experiments. Finally, a conclusion is drawn in section 5.

2 Related Work

At present, most of the web document clustering methods that can obtain good results only deal with the objects containing a lot of text information. Ramage et al. did many comparison experiments on whether tags could be used in the clustering process [2]. They prove that when tags are regarded as words with a higher weight, the clustering result can be improved. Perez-Tellez et al. proposed a method improving the clustering of blog posts through supplementing words from the text [3]. But in this method, the calculation of similarity is based on long text, tags only provide assistance. ZHANG et al. proposed a clustering method that did not use the contents of the blog post [4]. But it needs to cluster the tags before the calculation of similarity between blog posts. If the clustering results of the tags are not accurate, the accuracy of the similarity can not be assured either. Sun et al. analyzed the power law distribution of tags in the blogosphere [5]. They point out that only a small number of tags are used frequently. These tags are worth of attention in the analysis.

Begelman et al. in an earlier design of a tag clustering algorithm used co-occurrence frequency as the similarity between tags [6]. However, the co-occurrence frequency is influenced by the appearance frequency. It is not an accurate similarity measure when directly used. Simpson noticed the same problem during the process of tag clustering and used Jaccard coefficient to normalize the similarity between tags [7]. Xu et al. compared 8 commonly used similarity metrics of tags [8]. They prove that when tags are denoted as a vector according to the resources they tagged, the cosine value of the tag vector acquires the highest accuracy. Cattuto et al.'s tag clustering [9] and Boratto et al.'s tag semantic relevance analysis [10] also use cosine value as the similarity measure of tags. In fact, Jaccard coefficient and cosine value are both the co-occurrence frequency divided by some combination of the appearance frequency of the two tags. This will make a low co-occurrence frequency emerged by a very large appearance frequency and make the result of calculation very small though the two tags are of great relevance. Sun et al.'s study [11] on tag co-occurrence also shows that the cosine value will be affected by a tag with a broad concept and turn out to be inaccurate. Cui et al. point out that in the calculation of the similarity between tags, considering only co-occurrence frequency is not enough [12].

They proposed a method to calculate the similarity between the tags with no co-occurrence through link analysis between tags.

In this paper, we focus on the similarity calculation of tagged resources. The loss of efficiency in the similarity calculation of tags is not worth. The STAC measure just keeps the balance between accuracy and efficiency.

3 The STAC Measure

STAC is a similarity measure especially for tagged resources. It is based on the analysis of the co-occurrence information of tags and calculates the similarity between tagged resources through accumulating the similarity between tags. The accuracy of STAC can be assured while it is calculated with a high efficiency.

3.1 Notation and Definition

The tagging relations of tags and resources form a typical bipartite graph, as shown in Figure 1.

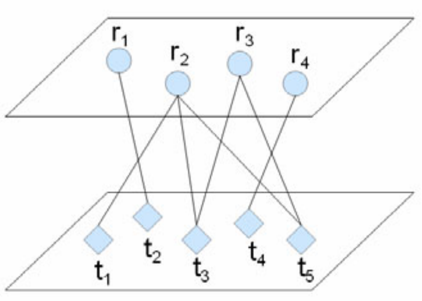


Fig. 1. The tagging relations between tags and resources

where R={r_1,r_2,r_3…} is the set of resources, T={t_1,t_2,t_3…} is the set of tags. If r_i is tagged by t_j, there is an edge between r_i and t_j in the graph.

In the remainder of this paper, we use:

- A to denote a tagging matrix describing the bipartite graph with size |R | × |T|, $A_{(i, j)}$=1 if there is an edge between r_i and t_j, $A_{(i, j)}$=0 otherwise;
- CoOccur(t_i, t_j) to denote the co-occurrence frequency of t_i and t_j, T_C to denote the co-occurrence matrix, where $T_{C(i, j)}$= CoOccur(t_i, t_j);
- CODMIN (t_i, t_j) to denote the similarity between tags, W_T to denote a similarity matrix describing the relation of tag nodes with size |T | × |T|, where $W_{T(i, j)}$= CODMIN (t_i, t_j);
- STAC(r_i, r_j) to denote the similarity between resources, W_R to denote a similarity matrix describing the relation of resource nodes with size |R | × |R|, where $W_{R(i, j)}$=STAC(r_i, r_j).

3.2 Similarity between Tags

Co-occurrence frequency is an important standard in the similarity measure of tags. If we already have matrix A, then

$$T_C = A^T \times A \tag{1}$$

However, it is not reliable to take co-occurrence frequency as the only basis. For example, in a group of tagging data “basketball” and “sports” co-occurred 653 times and reaches the highest co-occurrence frequency. The second highest frequency comes from “football” and “sports” and they only co-occurred 318 times. It is obvious that the similarity of the former two tags is not as twice high as the latter pair. The truth is the appearance frequency of “basketball” is almost twice as “football” (666:318). Therefore, the similarity between tags based on co-occurrence is often calculated by Jaccard coefficient and cosine value to eliminate the influence of appearance frequency of tags. But these two measures also have some problems. When one of the two tags has a very broad concept, its appearance frequency will be very high. This will emerge the co-occurrence frequency and make the similarity very low though the two tags are highly relevant. In STAC, we use CODMIN(t_i, t_j) to measure the similarity between tags.

$$CODMIN(t_i, t_j) = \frac{CoOccur(t_i, t_j)}{\min(Freq(t_i), Freq(t_j))} \tag{2}$$

Freq(t_i) is the appearance frequency of t_i. CoOccur(t_i, t_j) / Freq(t_i) can be regarded as the similarity between the two tags at the view of t_i and CoOccur(t_i, t_j) / Freq(t_j) at the view of t_j. Because similarity is bilateral, even calculated just at the view of one tag, the result can also reflect the similarity between the two tags. Moreover, as the smaller appearance frequency is more sensitive to the co-occurrence frequency, the reflection is more accurate. Meanwhile, this can assure the normalization of the similarity. We call Equation (2) CODMIN (Co-Occurrence frequency Divided by appearance frequency MINium) coefficient. Examples in Table 1 can further illustrate the superiority of CODMIN compared with Jaccard coefficient and cosine value.

Table 1. The comparison of Jaccard, Cosine and CODMIN

t_i	Freq(t_i)	t_j	Freq(t_j)	CoOccur(t_i, t_j)	Jaccard	Cosine	CODMIN
basketball	666	sports	1937	653	0.3288	0.5697	0.9805
football	327	sports	1937	316	0.1634	0.3996	0.9725
internet	165	it	1871	153	0.0813	0.274	0.9273
ferrari	38	auto	1910	38	0.0199	0.1411	1.0

CODMIN coefficient tends to raise the similarity between tags. For some pair of tags with low relevance, the result may be also high, due to the low appearance frequency of the tags. For example, “arc-shape” which appears only once co-occurs with “it”. The CODMIN of them is 1.0. Therefore, CODMIN does not suit for tag clustering. Even if there are such inaccurate tag similarity results, the calculation of STAC will not be affected, because the tags with low appearance frequency rarely participates the calculation of STAC. Those who play a key role in STAC are the tags with a high appearance frequency.

3.3 Calculation of STAC

After W_T is acquired through the calculation of CODMIN, the calculation of W_R can be regarded as a deduction from A and W_T to W_R. Considering the dimension of the three matrices, matrix W_{Sum} is calculated as follows

$$W_{Sum} = A \times W_T \times A^T \tag{3}$$

Equation (3) is a process of accumulating similarity through the comparison of tags. We call tag similarity $W_{T(i,j)}$ the *comparison result* of t_i and t_{ij}. For each element in matrix A × W_T, let the corresponding resource in its row be r_i and the corresponding tag in its column be t_j, then the value of this element is the accumulated comparison result of comparing t_j with every tag of r_i. For each element in matrix A × W_T × A^T, let the corresponding resource in its row be r_i and the corresponding resource in its column be r_j, the value of this element is the accumulated comparison result of comparing every tag of r_i with every tag of r_j. Finally, $W_{Sum(i,j)}$ represents the accumulated comparison result of resource r_i and r_j and that is a reflection of the similarity between the two resources. But it is obviously affected by the number of the tags of the resources. More tags mean more comparison times and larger accumulated result. The most effective solution is to divide $W_{Sum(i,j)}$ by the times of comparison. If W_T is replaced with matrix E whose elements are all 1, then we can get W_{Count}

$$W_{Count} = A \times E \times A^T \tag{4}$$

Equation (4) is a process of accumulating in which comparison results are all 1 and the result becomes the times of comparison. Let each element in matrix W_{Sum} be divided by the corresponding element of W_{Count}. We then acquire the more accurate similarity and the similarity Matrix W_R.

$$W_{R(i,j)} = W_{Sum(i,j)} / W_{Count(i,j)} \tag{5}$$

The result of the operation of matrices is the similarity matrix. After understanding the meaning of the operation, the formula to calculate the similarity between two resources can be easily acquired.

$$STAC(r_i, r_j) = \frac{\sum_{t_k \in rt_i, t_l \in rt_j} CODMIN(t_k, t_l)}{|rt_i| \cdot |rt_j|} \tag{6}$$

where rt_i, rt_j represent the sets of tags of r_i and r_j.

The meaning of STAC(r_i, r_j) is comparing every tag of r_i with every tag of r_j and calculating the average of the similarity between each pair of tags. It is also a normalized value between 0-1. STAC fully uses the information of each tag and avoids all kinds of adverse effects. It is a reliable similarity measure of tagged resources in the clustering process.

4 Experiments

According to the clustering method for traditional web documents, the resources are described as vectors by the VSM model. In this paper, we consider the condition in which only tags are used in the calculation of similarity, so the vectors only contain binary value, indicating whether the resource has a certain tag. Our experiments choose Euclidean distance and Jaccard coefficient to be the comparison similarity measures and choose the most intuitive partitioning method K-means and hierarchical method AGNES to be the clustering algorithm.

4.1 Data Set

This data set consists of blog posts crawled from http://blog.sina.com.cn, the biggest blog site in China. There is a category rank list in that site. The category a blog post belongs to can be regarded as a standard class label. This category information can be used to judge the clustering results in the experiment. There are 5000 blog posts from 10 categories in the data set, and each category has exactly 500 blog posts. All the blog posts have at least 5 tags and were delivered in 2009.

4.2 Criterion

We use *Purity* as the criterion of the clustering result. For each cluster C_i, *Main Category* is defined as the category that has the most blog posts and *MC* is the set of blog posts that belong to *Main Category* in this cluster. Define the purity of a cluster as

$$Purity(C_i) = \frac{|MC|}{|C_i|} \tag{7}$$

For the entire clustering results, define purity as

$$Purity = \sum_{i=1}^{k} \frac{|C_i|}{|D|} Purity(C_i) \tag{8}$$

where D is the set of all blog posts and k is the number of clusters.

4.3 Results

In the method of K-means, because the initial centers are randomly selected, in order to make a fair comparison, 10 groups of random centers are selected beforehand. For each group of centers, the similarity of blog posts is calculated using Euclidean distance, Jaccard coefficient and STAC respectively and then the K-means algorithm is executed. The purity distribution is shown in Figure 2.

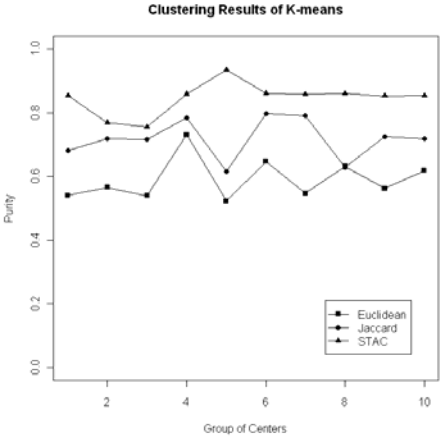


Fig. 2. Comparison results of K-means

In the method of AGNES, we choose MaxElementNum, the maximum of the number of elements in a cluster, as the terminating condition which is not relevant to the distribution of the similarity value. If the new cluster a combination step produce has more elements than MaxElementNum, this combination will be rejected. We choose 50-500 with an interval of 50, 10 MaxElementNum values to do the experiments. The purity distribution is shown in Figure 3.

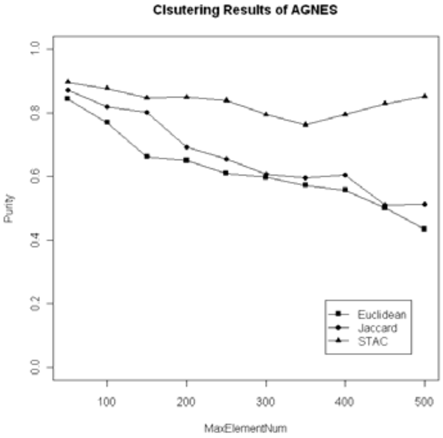


Fig. 3. Comparison results of AGNES

The experiments show that the clustering result of tagged resources using STAC is obviously better than Jaccard coefficient and Euclidean distance, no matter from best, worst or average purity.

5 Conclusion

In this paper, we propose STAC, a similarity measure especially for tagged resources. It solves the problem of calculating the similarity between tagged resources only through tags and therefore improves the clustering of tagged resources. Our study can benefit the analysis of tags, similarity calculation, clustering of tagged resources and other relevant research. Future work can try other deduction of the similarity matrix and improve the clustering algorithm of tagged resources.

Acknowledgments. This Work was supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2009AA012122) and Natural Science Foundation of LiaoNing Province (No.20072031).

References

1. Adam M.: Folksonomies - Cooperative Classification and Communication through Shared Metadata. In: Computer Mediated Communication - LIS590CMC (2004)
2. Daniel, R., Paul, H., Christopher, D.M., et al.: Clustering the Tagged Web. In: WSDM 2009, pp. 54–63 (2009)
3. Fernando, P., David, P., John, C., et al.: Improving the Clustering of Blogosphere with a Self-term Enriching Technique. In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNCS, vol. 5729, pp. 40–47. Springer, Heidelberg (2009)
4. Yin, Z., Kening, G., Bin, Z.: Clustering Blog Posts Using Tags and Relations in the Blogosphere. In: ICISE 2009, pp. 817–820 (2009)
5. Aixin, S., Maggy, A.S., Ying, L.: Blog Classification Using Tags: An Empirical Study. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 307–316. Springer, Heidelberg (2007)
6. Grigory, B., Philipp, K., Frank, S.: Automated Tag Clustering: Improving Search and Exploration in the Tag Space. In: WWW 2006, pp. 22–26 (2006)
7. Edwin, S.: Clustering Tags in Enterprise and Web Folksonomies. Technical report, HP Labs (2008)
8. Kaikuo, X., Yu, C., Yexi, J., et al.: A Comparative Study of Correlation Measurements for Searching Similar Tags. In: Tang, C., Ling, C.X., Zhou, X., Cercone, N.J., Li, X. (eds.) ADMA 2008. LNCS (LNAI), vol. 5139, pp. 709–716. Springer, Heidelberg (2008)
9. Ciro, C., Dominik, B., Andreas, H., et al.: Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems. In: LWA 2008, pp. 18–26 (2008)
10. Ludovico, B., Salvatore, C., Eloisa, V.: RATC: A Robust Automated Tag Clustering Technique. In: Di Noia, T., Buccafurri, F. (eds.) E-Commerce and Web Technologies. LNCS, vol. 5692, pp. 324–335. Springer, Heidelberg (2009)
11. Aixin, S., Anwitaman, D.: On Stability, Clarity, and Co-occurrence of Self-Tagging. In: WSDM 2009 (2009)
12. Jianwei, C., Pei, L., Hongyan, L., et al.: A Neighborhood Search Method for Link-Based Tag Clustering. In: ADMA 2009, pp. 91–103 (2009)

Advertising Keywords Extraction from Web Pages

Jianyi Liu, Cong Wang, Zhengyang Liu, and Wenbin Yao

School of Computer Science, Beijing University of Posts and Telecommunications
100876 Beijing, China
jianyilui@sohu.com

Abstract. A large and growing number of web pages display contextual advertising based on keywords automatically extracted from the text of the page, and it has been become a rapidly growing business in recent years. We describe a system that learns how to extract keywords from web pages for advertisement targeting. Firstly a text network for a single webpage is build, then PageRank is applied in the network to decide on the importance of a word, finally top-ranked words are selected as keywords of the webpage. The algorithm is tested on the corpus of blog pages, and the experiment result proves practical and effective.

Keywords: Keyword extraction, information extraction, advertising, PageRank.

1 Introduction

Content advertising systems are becoming an increasingly important part of the funding for free web services. The most notable online advertising platform examples include Google's AdSense program, MSN and Yahoo's Contextual Match product. These programs automatically find keywords on a webpage, and then show some dynamic relevant advertisements based on those keywords [1].

The general process of the typical content-targeted advertising systems is roughly as follows. When a user visits a webpage, such as a blog, a news page, or another source of information, the page URL will be sent to the Ad serving server, where the page is crawled and parsed. Prominent keywords or phrases will be extracted from the page and they will be used to find relevant Ad from the Ads database. Advertising appropriate to these keywords are displayed to the user. Typically, if a user clicks on the ad, the advertiser is charged a fee, most of which is given to the webpage owner, with a portion kept by the advertising service.

Picking appropriate keywords helps users in at least two ways. First, choosing appropriate keywords can lead to users seeing ads for products or services they would be interested in purchasing. Second, the better targeted the advertising, the more revenue that is earned by the webpage provider, and thus the more interesting the applications that can be supported. For instance, free blogging services and free email accounts with large amounts of storage are both enabled by good advertising systems [2].

Algorithms for keyword extraction can be classified into two broad categories: corpus dependent and independent approaches [3,4]. Corpus dependent approach

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 336–343, 2010.

requires a large stack of documents and predetermined keywords to build a prediction model. For example, Salton suggested the TF*IDF to capture the “weight” of a word in a collection of documents (using the words frequency of distribution). Other well-known proposed methods are Mutual Information (MI), by Church and Hanks; log-likelihood measure, by Dunning. Corpus independent approach directly sifts keywords from a single document without any previous or background information. For example, Matsuo proposed a keyword extraction algorithm based on the statistical analysis of a single document, starting from words-association measures of co-occurrence in a given context (i.e. the period).Generally it is accepted that corpus dependent approaches yield better performance. However, a prediction model is practically restricted to a single domain, thus the quality of extracted keywords from a new document of unknown domain is not always guaranteed. In this regard, corpus independent (or domain independent) approaches may find many practical applications. The best known programs for keyword extraction are Turney’s GenEx system [5], KEA and its variations [6-8].

This paper explains a text network based keyword extraction algorithm for a single webpage. Firstly a text network for a single webpage is build, then PageRank is applied in the network to decide on the importance of a word, finally top-ranked words are selected as keywords of the webpage. The algorithm needs no corpus and extracts keywords by analyzing the semantic structure of the whole text. The extracted keywords are semantically most important. The experiment result shows that the algorithm is effective and practical.

This paper is organized as follows. In Section 2, we describe the approach of text network based keyword extraction. Some experiment results are reported in Section 3. We conclude the paper in Section 4.

2 System Architecture

In this section, we introduce the general architecture of our keyword extraction system, which consist of the following four stages: preprocessor, text network building, PageRank computing and postprocessor. The keyword extraction based text network and PageRank proceeds as follows:

- Transform a webpage into text format, which preserve the structure and title information of the webpage.
- Tokenize and lemmatize the text, and annotate with pos tags.
- Identify text units that related to the text’s content and add them as nodes in the network, link the two nodes if they appear within a window size.
- Iterate computing the PageRank on the network until convergence.
- Sort the nodes according to their PageRank score and select top N nodes as the potential keywords of the text.
- Use the potential keywords’ pos tag information and position information in the text to select the keywords.

2.1 Preprocessor

The main purpose of the preprocessor is to transform an HTML document into an easy-to-process plain-text based document, while still maintaining important information. In particular, we want to preserve the blocks in the original HTML document, but remove the HTML tags. For example, text in the same table should be placed together without tags like <table>, <tr>, or <td>. The title information, which is the only human readable text in the HTML header, is an important source of useful information. The preprocessor first parses an HTML document, and returns blocks of text in the body, title information in the header. Because a keyword should not cross sentence boundaries, we apply a sentence splitter to separate text in the same block into various sentences.

2.2 Text Representation as a Network

The recent explosion of interest in networks had an immediate consequence the treatment of human language as a complex network. Actually some recent researches have shown that human language is clearly an example of a complex network [9-12]. A network is a system with interconnected components, where the components are called "nodes" and the connections "links". However, what is the "nodes" of language network? In fact, language exhibits highly intricate network structures at all levels (phonology, phonetics, morphology, syntax, semantics and pragmatics). In this paper, we will take words as the fundamental interacting units, not only because words is very common units in natural language processing, but also because it is relatively straightforward to obtain sufficient corpus data. What is the "links" of the language network that is how these words connect? According to the different relationships between words, we can build the following kinds of networks: Co-occurrence networks, dependency networks and semantic networks.

Considering the limit of efficiency and precision rate of semantic analysis and dependency analysis, our system apply the simplest and effective network— co-occurrence network. The text will be represented as a co-occurrence network, Graph= {V, E, W}, where V is a set of nodes, E is a collection of edges, W is the weights of edges.

- Node: In co-occurrence network, inter-words co-occurrence relation can partially reflect their syntactic and semantic relation in the text. Current researches mainly focus on this kind of network. However, not all words within the window size have relations. We find hubs of co-occurrence network for words with low semantic content but important grammatical functions (such as articles, auxiliaries, prepositions, etc). These functional words are the key elements that connect the text's structure, but they are not related to the text's content. Apparently, they should not be viewed as features of documents. So, these grammatical functions words are removed as stop-words. The remainder words in the text are viewed as the nodes of network, and a word only builds a node.
- Edge: Two nodes are connected if their corresponding lexical units co-occur within a window of maximum N words, where N can be set anywhere from 2 to 10 words. Their co-occurrence times are counted as the edge's weight.

2.3 PageRank on Text Network

As mentioned above, language network is an example of complex networks. Some graph-based ranking algorithms like Kleinberg's HITS algorithm or Google's PageRank, which have been successfully used in citation analysis, social networks, and the analysis of the link-structure of the World Wide Web, also can be used in language networks. A graph-based ranking algorithm is a way of deciding on the importance of a node within a graph, by taking into account global information recursively computed from the entire graph [13]. In the case of language network of a free text, a graph-based ranking algorithm can be applied to decide on the importance of a word. This paper selects Google's PageRank, which is widely used by search engines for ranking web pages based on the importance of the pages on the web.

The main idea is that: in a directed graph, when one vertex links to another one, it is casting a vote for that other vertex. The more votes one vertex gets, the more important this vertex is. PageRank also takes account the voter: the more important the voter is, the more important the vote itself is. In one word, the score associated with a vertex is determined based on the votes that are cast for it, and the score of the vertex casting these votes. So this is the definition:

Let G=(V,E) be a directed graph with the set of vertices V and set of edges E, when E is a subset of V×V. For a given vertex V_i, let In(Vi) be the set of vertices that point to it, and let Out(V_i) be the set of edges going out of vertex Vi. The PageRank score of vertex $\mathrm{V_i}$ is:

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{S(V_j)}{|Out(V_j)|} \tag{1}$$

d is a damping factor that can be set between 0 and 1,and usually set at 0.85 which is the value in this paper.

PageRank can be also applied on undirected weighted graph, in which case the out-degree of a vertex is equal to the in-degree of the vertex. The PageRank score of vertex V_i is:

$$S^*(V_i) = (1-d) + d * \sum_{j \in \mathrm{C}(V_i)} \frac{weight(E_{ij})S^*(V_j)}{|D(V_j)|} \tag{2}$$

C(V_i)is the set of edges connecting with V_j, weight(E_{ij})is the weight of edge E_{ij} connecting vertex V_i and V_j, and D(V_j) is the degree of V_j.

In our system, after the co-occurrence network is constructed (undirected weighted graph), the score associated with each vertex is set to an initial value of 1, and the ranking algorithm described in formula 2 is run on the co-occurrence network for several iterations until it converges – usually for 20-30 iterations [14].

Once a final score is obtained for each vertex in the network, vertices are sorted in reversed order of their score, and the top vertices in the ranking are retained for post-processing.

2.4 Postprocessor

After computing the PageRank on the co-occurrence network, a list of potential keywords ranked by PageRank score is generated. Postprocessor phase selects the keywords and reconstructs multi-word keywords using the following steps:

- Syntactic filter with nouns part of speech: The keyword of a text is usually a noun, so we only select the top nouns from the keywords list as potential keywords. While may be set to any fixed value, the number of keywords is usually ranging from 5 to 20.
- Reorder the top nouns list with title information: The title of a text, which is brief, informative, and usually describes contents accurately, contains the most important words. So we use the title information to reorder the top nouns list: if a word of the top nouns list occurs in the title, it will be reorder to the top.
- Reconstruct multi-word keywords: To avoid excessive growth of the graph size by adding all possible combinations of sequences consisting of more than one lexical unit (ngram), we consider only single words as candidates for addition to the graph. So we collapsed sequences of adjacent keywords into a multi-word keyword. For example, in the Fig.1, if both “文本表示” and “模型” are selected as potential keywords, since they are adjacent, they are collapsed into one single keyword “文本表示模型”.

3 Experiments

This section reports the experimental results comparing our system with several baseline systems, the comparisons between the variations of our system, including the impact of window size and keyword size. We first describe how the documents were obtained and annotated, as well as the performance measures.

3.1 Data and Evaluation Criteria

The first step was to obtain and label data, namely a set of web pages. Keywords are attached to the content of the text; however they are not defined in a consistent way. Therefore, we used author-based evaluation. We collected 1000 documents, which must have more than three tags as keywords, at random from the sohu blog (http://blog.sohu.com). Furthermore, blogs cover different domains such as news, sports, technology, entertainment, law etc. this will evaluate the performance of our algorithm on different domains.

Precision and coverage are selected as measures of performance. Precision is the result of the number of correct keywords divided by the number of extracted words. Coverage is the result of the number of correct keywords divided by the number of keywords the author has assigned.

3.2 Experimental Results

3.2.1 Effect of Iteration Times

The ranking algorithm described in formula 2 is run on the co-occurrence network for several iterations until it converges, but how to choose a suitable iteration times is a problem. This section will show the effect of iteration times to the algorithm. Here we choose 10, 20 and 30 as iteration times. The default window size is 10 and the number of extracted keywords is 5,6,7,10,15 separately. Table 1 shows the result of each situation.

Table 1. Effect of iteration times

Times Number	10		20		30	
	Precision	Coverage	Precision	Coverage	Precision	Coverage
5	68.9	62.9	69.1	63	69.1	63.1
6	61.2	67.2	61.6	67.3	61.6	67.3
7	54.9	69.7	55.1	70	55.1	70
10	40.4	72.4	40.8	73.8	40.8	73.8
15	28.1	75.7	28.4	76.3	28.4	76.3

From Table 1, we can see that there is little change in precision and coverage between 10 and 20 iteration times. Above 20 iteration times, the results are more stable and almost the same. Table 1 shows the PageRank algorithm achieves convergence after 20-30 iterations.

3.2.2 Effect of Window Size

Text network is basic structure in our algorithm, so the way of building text network is especially important. Among many factors, the most important one is the window size, which controls how to build edges of between vertices. The window size is maximum words distance within which there can be an edge in the graph. According to different window size, one document can be presented as different text networks. This will make a remarkable effect on PageRank computation results. Here we choose 2, 3, 5, 10, and 15 as window size. The default iteration time is 20 and the number of extracted keywords is 5,6,7,10,15 separately. Table 2 shows the result of each situation.

From Table 2 we can see that the experimental result obviously rise with window size changing from 2-10. But when the window changes from 10 to 15 the results only have a tiny increase and the result even decrease after the window size reach 15. A bigger window size will cost more time to run the algorithm, so we must find a balance between quality and effective. Generally, we set default window size as 10.

Table 2. Effect of Window size

Num Window	5		6		7		10		15	
	P	C	P	C	P	C	P	C	P	C
2	55.8	50.9	49.9	54.6	44.9	57.1	34.2	62	24.8	66.6
3	60.2	54.8	54.2	59.1	48.6	61.8	36.9	66.7	26.1	70
5	64.4	58.7	57.7	63	52.2	66.4	39	70.6	27.5	73.7
10	**69.1**	63.1	61.6	67.3	55.1	70	40.8	73.8	28.4	76.3
15	68.3	62.3	61.3	67.1	54.9	69.8	40.9	73.9	28.5	**76.4**

3.2.3 Comparison Result

A most popular algorithm for keyword extraction is the tf*idf measure, which extracts key words that appear frequently in a document, while seldom in the remainder documents of the corpus [15]. As a comparison, tf*idf is also used to extract keywords. Elimination of stop words and stemming are processed ahead and 10 most frequent words of the text are extracted as keywords.

Table 3. Comprarison results

	Coverage	Precision
tf*idf	51.2	26.7
PageRank on Text network	73.8	40.8

Results are shown in Table 3. tf*idf ignores semantic structure of a document, transforms the document form a string of characters into a bag of words, and assumes the words is independent. While text network represents a text as semantic network, decides the importance of nodes by taking into account global information recursively computed from the entire network, rather than relying only on local node-specific information, and regards those top "n" important nodes as keywords. Therefore, language network can detect some "hidden" keywords even if they do not appear frequently, which are not been extracted by tf*idf.

4 Conclusions

In this paper, we introduced a text network based advertising keyword extraction from web pages. Keyword extraction is an important technology in many areas of document processing. Firstly a text network for a single web page is build, then PageRank is applied in the network to decide on the importance of a word, finally top-ranked words are selected as keywords of the document. The algorithm is tested on the corpus of sohu blog, and the experiment result proves practical and effective.

Acknowledgements. This study is supported by Beijing Natural Science Foundation of (4092029) and the Fundamental Research Funds for the Central Universities (2009RC0217).

We also thank the anonymous reviewers for their constructive comments.

References

1. Jin, X., Li, Y., Mah, T., Tong, J.: Sensitive Webpage Classification for Content Advertising. In: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising, San Jose, California, August 12-12, pp. 28–33 (2007)
2. Yih, W.-T., Goodman, J., Carvalho, V.R.: Finding Advertising Keywords on Web Pages. In: Proceedings of WWW 2006 (2006)

3. Yang, W., Li, X.: Chinese keyword extraction based on max-duplicated strings of the documents. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 439–440 (2002)
4. Panunzi, A., Fabbri, M., Moneglia, M.: Keyword Extraction in Open-Domain Multilingual Textual Resources. In: First International Conference on Automated Production of Cross Media Content for Multi-Channel Distribution, pp. 253–256 (2005)
5. Turney, P.D.: Learning algorithms for keyphrase extraction. Information Retrieval 2(4), 303–336 (2000)
6. Frank, E., Paynter, G.W., Witten, I.H., Gutwin, C., Nevill-Manning, C.G.: Domain-specific keyphrase extraction. In: Proc. of IJCAI 1999, pp. 668–673 (1999)
7. Mitchell, T.: Tutorial on machine learning over natural language documents (1997), `http://www.cs.cmu.edu/~tom/text-learning.ps`
8. Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge. In: Proc. of EMNLP 2003, pp. 216–223 (2003)
9. Ferreri Cancho, R., Solé, R.V.: The small-world of human language. In: Proceedings of the Royal Society of London, pp. 2261–2266 (2001)
10. Dorogovtsev, S.N., Mendes, J.F.F.: Language as an evolving word web. In: Proc. Royal Soc. London, pp. 2603–2606 (2001)
11. Sole, R.V., Corominas, B., Valverde, S., Steels, L.: Language Networks: their structure, function and evolution, Trends in Cognitive Sciences (2005)
12. Luoxia, W., Yong, L., Wei, L., et al.: 3-degree Separation and Small World Effect of Chinese Character Network. Chinese Science Bulletin 49(24), 2615–2616 (2004)
13. Wang, J., Liu, J., Wang, C.: Keyword Extraction Based on PageRank. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 738–746. Springer, Heidelberg (2007)
14. Mihalcea, R., Tarau, P.: TextRank: Bringing order into texts. In: Proceedings of EMNLP, pp. 404–411 (2004)
15. Matsuo, Y., Ishizuka, M.: Keyword Extraction from a Single Document using Word Cooccurrence Statistical Information. In: Proceedings of the 16th International FLAIRS Conference, St. Augustine, Floridam (2003

Automatic Topic Detection with an Incremental Clustering Algorithm

Xiaoming Zhang[1] and Zhoujun Li[2]

[1,2] School of computer, Beihang University, Beijng, China
yolixs@163.com, Zhoujun.li@263.net

Abstract. At present, most of the topic detection approaches are not accurate and efficient enough. In this paper, we proposed a new topic detection method (TPIC) based on an incremental clustering algorithm. It employs a self-refinement process of discriminative feature identification and a term reweighting algorithm to accurately cluster the given documents which discuss the same topic. To be efficient, the "aging" nature of topics is used to precluster stories. To automatically detect the true number of topics, Bayesian Information Criterion (BIC) is used to estimate the true number of topics. Experimental results on Linguistic Data Consortium (LDC) datasets TDT4 show that the proposed method can improve both the efficiency and accuracy, compared to other methods.

Keywords: TDT, Topic Detection, incremental clustering, term reweighting.

1 Introduction

Topic detection is a main task of TDT, and it is the problem of identifying stories in several continuous news streams that pertain to new or previously unidentified events [3]. In the other side, topic detection is a problem of grouping all stories as they arrive, based on the topics they discuss. This task differs from standard document clustering, where the objective is to group topically related documents into clusters that capture general categories or topics and the number of clusters is defined by user. We define a topic over a corpus to be a set of documents that share a consistent theme or concept. Two documents can lie in the same topic yet still cover different specific issues, e.g. a news article on a forest fire and the other one that reports an earthquake are both members of the topic "Natural Disasters". It is possible to imagine any number of of equally valid topic boundaries for a particular dataset, For topic Detection, we aim for vlusters that reflect the full narrative of an event as it grows and develops over time. Unlike a set of topic, there are a finite number of valid events that could take place for a collection of TDT documents.

In topic detection, another important factor that affects the performance is the judgment of the similar topics, as there are a great many words which are the same in these different but similar reports and they are easy to lead a miscarriage of thematic justice. Although, to some degree, it can help to differentiate similar topics by named entity [5], the number of named entity is limited in news reports. What's more, only by the named entity, it may cause damage to the thematic framework as many key

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 344–351, 2010.

words which describe the contents are ignored. As a result, the performance of topic detection can't be improved.

In this paper we propose an automatic topic detection method (TPIC) based on an incremental clustering algorithm. Each topic is represented by a number of sub-topics, and similarity between topics is measured by the smallest similarity between the sub-topics of the cluster pair. To accurately cluster the given document corpus, we employ a discriminative feature set to represent each document, and reweight each feature by using document distribution and other information. From the initial clustering, the operation that refines the discriminative feature set is iteratively applied in the procedure of clustering. On the other hand, Bayesian Information Criterion (BIC) is use to estimate the true number of topics. We use BIC to determine whether two topics can be merged to one cluster. Experiment results indicate that this topic detection is more accurate and can estimate the true number of topics.

2 Related Works

There have many works which relate to topic detection. The language modeling approach [6,7] assumes a multi-stage generative process where semantically meaningful modalities such as topics or authors are chosen as an intermediate step, and then the words are drawn from the multinomial distribution conditioned on these modalities. Graph properties are used to study community structures by [12-17]. As a distance metric [12] which uses the similarity of citation patterns, paper [13] use the notion that nodes have more links to the members of the same community than to other nodes. Paper [14] introduces the concept of edge betweenness, and paper [15] uses the measures from bibliometry and graph theory. Some papers in this group combine the information from text as well. Paper[16] extracts storylines for a query by identifying densely connected bipartite from the document-term graph of the search results. Paper[17] improve the document categorization performance by starting from a text-based categorization result and then iteratively relabel the documents to further satisfy the constraints imposed by the link proximity relation.

3 Topic Detection Based on Incremental Clustering

In this section, we propose an efficient topic detection model based on an incremental clustering algorithm, in which features are refined in the procedure of clustering.

3.1 Terms Weighting

TF-IDF is the most prevalent terms weighting method in information retrieval systems. However, the TF-IDF method can't weight terms of some classes properly. For example, terms that occurs frequently in one topic but infrequently in other topics, and terms with low document frequency but appears in different topics. Besides, the *TF-IDF* rarely considers the importance of document weight and document distribution. In fact, documents are also important in discrimination of terms. The main assumption behind document weighting is as following: the more information a document gives to terms the more effect it gives to latent variable, and the less information a document gives to terms the less effect it gives to latent variable.

To address above problems, we propose that term weight is constituted of following parts.

$$W(i,j) = LT(i,j) \times GT(i) \times GD(j) \tag{1}$$

The notation used in the follow equations is defined as:
tf_{ij}: the frequency of term *i* in document *j*.
dl_j: the number of documents that contain term *j*.
df_{ci}: the number of documents containing term *i* within cluster *c*.
gf_i: the frequency of term *i* in document collection.
sg_f: the sum frequency of all terms in document collection.
N_c: the number of documents in cluster c.
N_t: the total number of documents in collection.

We replace the *TF* in *TF-IDF* with *LT*. In this equation, the length of document is used, because that a feature is likely to accrue more frequently in longer document than in shorter document.

$$LT(i,j) = \frac{\log(tf_{ij}+1)}{\log dl_j + 1} \tag{2}$$

Entropy theory is used to set *GT(i)* and *GD(j)*, and it replaces IDF with following.

$$GT(i) = \frac{H(d) - H(d \mid t_i)}{H(d)} = 1 - \frac{H(d \mid t_i)}{H(d)} \tag{3}$$

$$H(d \mid t_i) = -\sum p(j \mid i) \log p(j \mid i) \quad , \quad p(j \mid i) = \frac{tf_{ij}}{gf_i} \tag{4}$$

$$H(d) = \log(N_t) \tag{5}$$

$$GD(j) = \frac{H(t) - H(t \mid d_j)}{H(t)} = 1 - \frac{H(t \mid d_j)}{H(t)} \tag{6}$$

$$H(t) = -\sum p(t) \times \log p(t) = -\sum_{i=1}^{n} \frac{gf_i}{sgf} \times \log \frac{gf_i}{sgf} \tag{7}$$

$$H(t \mid d_j) = -\sum p(i \mid j) \log p(i \mid j) \quad , \quad p(i \mid j) = \frac{tf_{ij}}{dl_j} \tag{8}$$

In the equation of *GT(i)*, the frequency of document in document j and in the collection are all considered, and it can reduce the affection of document length. The equation of *GD(j)* mainly computer the importance that the document to the term. If a document is important, it influences the terms greatly.

3.2 Refining of Feature Set

In most clustering method, they treat all the features equally. Thus, we discover a group of discriminative features from the initial clustering result, and then iteratively refine the discriminative features and cluster the document set using this discriminative feature set. To determine whether a feature is discriminative or not, we define the following discriminative feature metric *KD(i)*:

$$KD(i) = KL(P_{ci} \parallel P_{ti}) \tag{9}$$

$$P_{ci} = \frac{df_{ci}}{N_c}, P_{ti} = \frac{df_i}{N_t} \tag{10}$$

$$KL(P \parallel Q) = \sum p(x) \log \frac{p(x)}{q(x)} \tag{11}$$

where P_{ci} denotes the ratio between the number of documents that contain feature *i* and the number of the total documents in cluster *c*, and P_{ti} denotes the ratio between the numbers of document that contain feature i and the number of total document. Obviously, *KD(i)* is used to enhance the weights of terms that occur frequently in a topic, and infrequently in other topics. In other words, the more discriminative the feature f_i, the larger value the metric *KD(i)* takes. In our real implementation, the weight of the top-50% features that have the greatest *KD(i)* value is multiplied by *(1+KD(i))*.

$$W(i,j)^{re} = W(i,j) \bullet (1 + KD(i)) \tag{12}$$

3.3 BIC Score

Using model selection techniques has been applied in many clustering algorithms to determine the true number of clusters. The problem of model selection is to choose the best one among a set of candidate models. Let $\{x_1, \ldots, x_n\}$ be a set of input data *D*, where each $x_i \in R^d$, and *D* can be partitioned into disjoint subset $C_1, \ldots, C_k$. In this paper, *k* is the number of topics. The *BIC* of the model M_i is defined as:

$$BIC(M_i) = \hat{l}_i(D) - \frac{p_i}{2} * \log n \tag{13}$$

where $\tilde{l}_i(D)$ is the log-likelihood of the data according to the model M_i and taken at the maximum likelihood point, and p_i is the number of independent parameters in M_i. The probability that a data point x_i belongs to a cluster C_j is defined as the product of the probability of observing C_j and the multivariate normal density function of x_i:

$$\hat{P}(x_i) = \frac{n_j}{n} \bullet \frac{1}{\sqrt{2\pi \hat{\sigma}^d}} \exp(-\frac{1}{2\hat{\sigma}^2} \parallel x_i - \mu_j \parallel^2) \tag{14}$$

where n_j is the number of points in the cluster C_j , and $\overset{\wedge}{\sigma}{}^2$ is the maximum likelihood estimate (MLE) of the variance defined by:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_i (x_i - \mu_j)^2 \tag{15}$$

Thus the maximum log-likelihood of the data in cluster C_j can be calculated as:

$$\hat{l}(C_j) = \log \prod_{i \in C_j} \hat{P}(x_i) == -\frac{n_j}{2}\log(2\pi) - \frac{n_j \bullet d}{2}\log(\hat{\sigma}^2) - \frac{n_j - k}{2} + n_j \log n_j - n_j \log n \tag{16}$$

Finally, the BIC can be written as follows:

$$BIC(M_i) = \sum_{j=1}^{k} \hat{l}(C_j) - \frac{p_i}{2} \bullet \log n \tag{17}$$

Given a set of candidate models, the model with the highest *BIC* score is selected. We calculate the *BIC* locally when the clustering algorithm performs the merging test in each cluster pair. If the *BIC* score of the new cluster structure is less than the current *BIC* score, we do not merge the two clusters.

3.4 Incremental Topic Detection

For each cluster, *c* subcluters are stored. A subtopic is the mean of all the points in this subcluster. The subtopics are determined by first choosing *c* well scattered points within the cluster, and these points usually are the farthest points from the center of cluster. Then use these points to construct sub-clusters in which all the points are their nearest neighbors in the cluster, and the mean of each sub-cluster represents a subtopic. The distance between two clusters is then the distance between the closest pair of subtopics - one belonging to each of the two clusters. Thus, only the subtopics of a cluster are used to compute its distance to other clusters.

For each cluster *X*, *X*.mean and *X*.subpoints store the mean of the points in the cluster and the set of *c* scattered points for the cluster, respectively. For a pair of points *X, Y*, *dist(X,Y)* denotes the distance between the points. This distance could be any of metrics, we use cosine metric in this paper. Alternatively, numeric similarity functions can also be used. The distance between two clusters *X* and *Y* can then be defined as

```
Procedure Detection
Input documents D;
Output topics x1...xk;
1.   { T=build_tree(D);
2.   ST=heap_pair(T)
3.   while(ST!=Φ){
4.   X,Y=extract_closest_pair(ST);
5.   Z= merge(X,Y);
6.   if (BIC(Z>BIC(X,Y))){
7.   update T;
8.   update ST;
9.   refine feature set;
10.  }}}
```

Fig. 1. Procedure of topic detection

In figure 1, a stack is used to store the cluster pairs in step 3, and it is arranged in the increasing order of the distance between the cluster pair. When the stack is empty, the procedure of topic detection is completed. In step 5, a pair of closest clusters is extracted from the heap, and then they were merged (explained in figure 2) in step 6. *BIC* is used to determine whether two clusters can be merged in step 7. If the two clusters can be merged, then clusters tree (*T*) and heap (*ST*) are updated accordingly in step 8 and 9. In the tree updating process, a new cluster *Z* is inserted, and cluster *X* and *Y* are deleted. Afterward, *Z*'s closest count part is found, and the heap is also updated. The merge procedure in figure 2.

As the input size *n* grows, the computation that needs to be performed by the detection algorithm could end up being fairly substantial due to the $O(n^2 log\ n)$ time complexity. As it is known, topics have "aging" nature, which means that an old inactive topic less likely attracts new stories than recently active events. Therefore, stories that happen in closest time are more likely to be of the same topic. The basic idea of pre-clustering is that all the stories are ordered by their published time, and then partition all the stories into partitions by measure of periods.

```
procedure merge(X, Y)
{
Z:=X ∪ Y;
Z.mean= $\frac{|X|X.mean+|Y|Y.mean}{|X|+|Y|}$;
tempest=Φ;
for( i=1 to c do){
maxDist=0;
for each point p in X.subpoints ∪ Y.subpoints do{
if i=1;
    Dist=dist(p, Z.maen)
else
    Dist=min{dist(p,q): q∈tempSet}
  if(Dist≥maxDist){
     maxDist=Dist;
     maxpoint=p;
    }}
    tmpset=tmpset∪{maxpoint}
  }
  Z.subpoints=tmpset;
  sub_clusters=build_subcluster(tmpset);
  foreach point p in Z do
     q=p's nearest point in tmpset
     allocate p to q's sub-cluster;
     computing sub_mean of each sub-cluster; }
return Z;
```

Fig. 2. Procedure of merging clustering

4 Experiments

In the evaluation, we used the standard corpora TDT-4 from the NIST TDT corpora. Only English documents are chosen for evaluation. We use the precision, recall and F1 that are used to evaluate the performance of clustering generally to evaluate the performance of topic detection.

To test our approach, we implement three topic detection systems:

System 1 (*K-means*), in which all the stories on hand are clustered by *k-means*, and each cluster represent a topic.

System 2 (*CMU*), this method divides the incoming data stream into sequential buckets, clustering by cosine similarity within each bucket, and then merging clusters among buckets [7].

System 3 (*TPIC*), it is implemented based our approach.

In the first experiment we compare the recall, precision, miss and F1 of different systems. Figure 3 summarize the results of the three Systems. All these systems conducted multiple runs with different parameter settings; here we present the best result for each system with respect to the F1 measure. As shown, the results of the evaluations demonstrate that the proposed topic detection model outperforms other models greatly. Because in *TPIC*, there has subtopic points and feature refining process, which can alleviates the shortcomings of centroid-based approach of other systems.

	K-means	CMU	TPIC
Recall(%)	50	62	80
Precision(%)	48	82	84
Miss(%)	50	35	12
False Alarm(%)	0.25	0.13	0.155
F_1	0.59	0.70	0.84

Fig. 3. Topic detection results

In the second experiment, we mainly compare the time consumed by the three systems. For the experiment data with different number of stories, we select these stories from corpora randomly. Figure 4 shows the execution time of different system. The result indicates that *TPIC* has the least execution time compare to other two systems. This is because that the pre-clustering operation in TPIC can greatly reduce the comparing time of cluster pair.

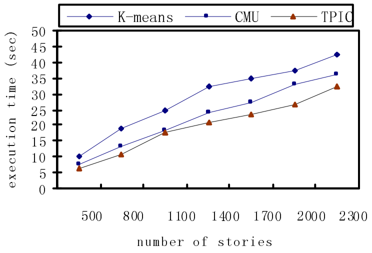


Fig. 4. Comparison of execution time

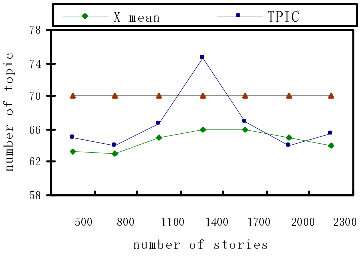


Fig. 5. Comparison between X-means and TPIC

Another experiment is used to test how good TPIC is at revealing the true number of topics or clusters. In this experiment, we compare TPIC with X-means which also use BIC to estimate the true number of cluster based on K-means [20]. Detailed results for 70 topics case are in figure 5. It shows that TPIC outputs a configuration which is within 10% from the true number of topics. It also show that TPIC is not worse than X-means, and is better than X-means in some times, and the performance of local BIC is almost as well as it of global BIC.

5 Conclusion

The challenge of topic detection is to cluster large number of stories based on the topics they discuss. There have been many clustering algorithms, but they can't be used to topic detection directly because news stories have their own characters such as viewpoint, "aging" and so on. In this paper, we proposed a new topic detection model (*TPIC*) based on an incremental clustering algorithm. The incremental clustering algorithm can estimate the true number of cluster, and use sub-topic points and refining feature set to improve the performance of topic detection. Pre-clustering operation using "age" feature of stories is used to reduce the execution time of clustering procedure. We compare the performance of systems based on *TPIC*, *K-means* and *CMU* using experiments. Experiments results show that *TPIC* has a higher performance and less execution time the other two models.

References

1. Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., Liu, X.: Learning approaches for detecting and tracking news events. IEEE Intelligent Systems: Special Issue on Application of Intelligent Information Retrieval 14(4), 32–43 (1999)
2. Wang, Z.M., Zhou, X.H.: A Topic Detection Method Based on Bicharacteristic Vectors. In: Proceeding International Conference on Networks Security. Wireless Communications and Trusted Computing (2009)
3. Jo, Y.K., Lagoze, C., Lee Giles, C.: Detecting research topics via the correlation between graphs and texts, pp. 370–379 (2007)
4. Griffiths, T.I., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review 114(2), 211–244 (2007)
5. Hopcroft, J., Khan, O., Kulis, B., Selman, B.: Natural communities in large linked networks. In: Proceedings of SIGKDD (2003)
6. Ino, H., Kudo, M., Nakamura, A.: Partitioning of web graphs by community topology. In: Proceedings of WWW (2005)
7. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. arXiv:cond-mat/0308217 (2003)
8. Newman, M.: Scientific collaboration networks. i. network construction and fundamental results. Physical Review E 64 (2001)
9. Kumar, R., Mahadevan, U., Sivakumar, D.: A graph-theoretic approach to extract storylines from search results. In: Proceedings of SIGKDD (2004)
10. Angelova, R., Weikum, G.: Graph-based text classification: Learn from your neighbors. In: Proceedings of SIGIR (2006)
11. Pelleg, D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Cluster. In: Proc. of the 17th International Conf. on Machine Learning, pp. 727–734 (2000)

Three New Feature Weighting Methods for Text Categorization

Wei Xue and Xinshun Xu*

School of Computer Science and Technology, Shandong University,
Jinan 250101, China
xueweicode@gmail.com, xuxinshun@sdu.edu.cn

Abstract. Feature weighting is an important phase of text categorization, which computes the feature weight for each feature of documents. This paper proposes three new feature weighting methods for text categorization. In the first and second proposed methods, traditional feature weighting method *tf*×*idf* is combined with "one-side" feature selection metrics (i.e. odds ratio, correlation coefficient) in a moderate manner, and positive and negative features are weighted separately. *tf*×*idf*+*CC* and *tf*×*idf*+*OR* are used to calculate the feature weights. In the third method, *tf* is combined with feature entropy, which is effective and concise. The feature entropy measures the diversity of feature's document frequency in different categories. The experimental results on Reuters-21578 corpus show that the proposed methods outperform several state-of-the-art feature weighting methods, such as *tf*×*idf*, *tf*×*CHI*, and *tf*×*OR*.

Keywords: feature weight, feature selection, text categorization.

1 Introduction

In recent years, the volume of text documents available on the internet has grown rapidly. Text categorization plays an important role in indexing, organizing and categorizing these eclectic documents. Automatic text categorization is a process of assigning predefined category labels to text documents. A number of machine learning techniques have been applied to text categorization, for instance, Support Vector Machine (SVM) [1], k-Nearest Neighbors (KNN) [2], Maximum Entropy [3], and AdaBoost [4].

These machine learning techniques can't process natural language documents directly, so documents usually are considered as bags of words (BOW). In the BOW model, the word ordering and text structure are ignored. Vector Space Model (VSM) is a widespread method for representing documents. In this model, a document d_i is represented by a feature vector $d_i = (w_{i1}, w_{i2}, \ldots, w_{in})$, where w_{ij} is the weight of feature (also called term) t_j in d_i. The wildly used feature weighting method is *tf*×*idf*, where *tf* (term frequency) represents the importance of a feature in a document, and *idf* (inverse document frequency) represents the discrimination of a feature for all

* Corresponding author.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 352–359, 2010.

documents. *tf*×*idf* is borrowed from Information Retrieval (IR), and *idf* was introduced to prevent most documents are retrieved. However, both *tf* and *idf* ignore the fact that features have different discriminations for distinct category labels.

Local feature selection is a process of selecting a subset of the original feature set, and most feature selection metrics rank features according to their discriminating capabilities for each category. Many feature selection methods have been employed in text categorization, such as chi-square (CHI), odds ratio (OR), correlation coefficient (CC), information gains (IG), mutual information (MI), GSS coefficient (GSS) [5], [6]. Most of them can be divided into two classes: "one-side" metrics (e.g., OR, CC and GSS) and "two-side" metrics (e.g., CHI) [7], [8]. The key property of "one-side" metrics is that they select all the positive features that are indicative of membership in a category before they consider the negative features that are indicative of non-membership. The positive features are assigned positive scores and the negative features are assigned negative scores. On the other hand, "two-side" metrics don't differentiate positive features from negative features. All features are assigned positive scores, and positive features are mixed with negative features.

In order to embody the feature weight with category discriminating capability, Debole [9] proposed originally the supervised term weighting (STW). In STW model, feature selection functions are introduced into the feature weights. Deng et al. [10] replaced *idf* with feature selection metrics, and *tf*×*CHI* and *tf*×*OR* are used. However, although OR is a "one-side" feature selection metric, when *tf*×*OR* is used for SVM classifiers with a linear kernel, it equals to *tf*×|*OR*|. As a result, *tf*×*OR* has the same effect as "two-side" metrics. Both *tf*×*CHI* and *tf*×*OR*, enlarge the contribution of positive features as well as negative features.

In this paper, three new feature weighting methods are proposed. First, *tf*×*idf* is combined with "one-side" metrics, and positive features and negative features are weighted separately. The proposed methods enlarge the positive feature's weight, and don't change other features. The first and second methods are *tf*×*idf*+*OR* and *tf*×*idf*+*CC* respectively. In the third method, a new metric called feature entropy (FE) is introduced, which contains the membership of documents and more effective and concise than most feature selection metrics. Experimental results on Reuters-21578 corpus show that the new methods have better performances than *tf*×*OR*, *tf*×*CHI* and *tf*×*idf*.

The remainder of the paper is organized as follows. Section 2 describes related works. Section 3 describes the proposed feature weighting methods. Experimental results and discussions are presented in section 4. Finally, section 5 concludes this work and gives the future work.

2 Related Work

Several feature weighting methods for text categorization have been studied in previous works. Debole [9] compared *tf*×*idf*, *tf*×*CHI*, *tf*×*GR* and *tf*×*IG* on Reuters-21578, and SVM-light was used for classifying. The result of ten most frequent categories shows that *tf*×*CHI* and *tf*×*GR* have better performance than *tf*×*idf* on macro-averaging

F_1, but on micro-averaging F$_1$, *tf*×*idf* is the best. Deng et al. [10] also reported experimental results of *tf*×*idf*, *tf*×*CHI* and *tf*×*OR*, and their result on Reuters-21578 (all categories) reveals that *tf*×*CHI* is the best, with 1 percent higher than *tf*×*idf* on micro-averaging F_1. Both Debole [9] and Deng [10] employed feature selection in their experiments.

3 The Proposed Methods

The traditional feature weighting approach is *tf*×*idf*, where *tf* denotes the frequency of a feature in a document, and *idf* denotes the number of documents that a feature occurred in. The *tf*×*idf* is defined as [11]:

$$tfidf(t_i,d_j)=tf_{ij}\times idf_i=P(t_i,d_j)\times\log\frac{N}{P(t_i)}\,. \tag{1}$$

where $P(t_i,d_j)$ represents the frequency of a term t_i occurred in the document d_j, N is the total number of documents, and $P(t_i)$ denotes the number of documents that contained t_i.

tf represents the importance of a feature in each document, and *idf* punishes terms that occured in most documents. All of them ignore the discrimination of a feature for each category. To solve the shortage of *tf*×*idf*, we construct new feature weighting functions: *tf*×*idf*+*OR* and *tf*×*idf*+*CC*. As Sebastiani [6] reported that OR and CC perform better than CHI, IG and MI. OR and CC are used in the new functions, which are “one-side” feature selection metrics and distinguish positive from negative features.

idf only considers the document frequency of terms in the whole collection. To integrate document frequency and category information, a new metric is introduced and combined with *tf*.

The detailed descriptions of the proposed method are given in the following subsections.

3.1 tf×idf+OR

Odds ratio (OR) measures the ratio of the frequency of a term occurred in relative documents to the frequency of this term occurred in non-relative documents. The basic idea is that the distribution of features on relevant documents is different from the distribution of features on non-relevant documents [8]. OR was originally proposed by Van Rijsbergen [12] for Information Retrieval, and first used in text categorization by Mladenic [13]. It’s defined as follows:

$$OR(t_i,c_j)=\log\frac{P(t_i\mid c_j)(1-P(t_i\mid\overline{c_j}))}{(1-P(t_i\mid c_j))P(t_i\mid\overline{c_j})}\,. \tag{2}$$

where $P(t_i|c_j)$ is the conditional probability of that t$_i$ appears in documents labeled c_j, and $P(t_i/\overline{c_j})$is the conditional probability of that t$_i$ appears in documents that not labeled c_j.

Equation (2) can be estimated by using:

$$OR(t_i, c_j) \approx \log \frac{AD}{CB}. \tag{3}$$

where A (B) denotes the number of documents that t_i occurs in and (not) labeled c_j; C (D) denotes the number of documents that t_i does not occur in and (not) labeled c_j.

Deng et al. [10] used *tf*×*OR* to calculate the feature weight. But the OR scores of positive features are plus quantity, and the scores of negative features are minus. For a positive feature t_m and a negative t_n, if $|OR(t_m,c_j)| = |OR(t_n,c_j)|$, then the contributions of these two features to the inner product of two documents will be the same when kernels are used in classifiers. To distinguish positive features from negative features, we construct a new weighting function *tf*×*idf*+*OR*, defined as follows:

$$w_{tf\times idf+OR} = \begin{cases} tfidf + \alpha \times OR & \text{if } tfidf > 0 \ \&\ OR > 0 \\ tfidf & \text{otherwise} \end{cases}. \tag{4}$$

where *d* represents a document, *tfidf* denotes the *tf*×*idf* weight of term t_i in document *d*, and *OR* defined by equation (3) denotes the OR score of t_i in the category that d belongs to. α is a constant. In this function, only positive features' weights are larger than their *tf*×*idf* values. Negative features' weights are equal to their *tf*×*idf* values. As a result, positive features' contributions to document's similarity are larger than negative feature's.

3.2 tf×idf+CC

Correlation coefficient (CC) is a variant of chi-square (CHI) metric, where $CC^2 = CHI$. So CC can be viewed as a "one-side" CHI metric [8]. The CC of a term t_i for a category c_j can be defined as [14]:

$$CC(t_i, c_j) = \frac{\sqrt{N}(AD - CB)}{\sqrt{(A+C)(B+D)(A+B)(C+D)}}. \tag{5}$$

where *N* is the total number of documents. Then the feature weighting function *tf*×*idf*+*CC* can be defined as follows:

$$w_{tf\times idf+CC} = \begin{cases} tfidf + \alpha \times CC & \text{if } tfidf > 0 \ \&\ CC > 0 \\ tfidf & \text{otherwise} \end{cases}. \tag{6}$$

where *d* represents a document, *tfidf* denotes the *tf*×*idf* weight of term t_i in document *d*, and *CC* defined by equation (5) denotes the CC score of t_i in the category that *d* belongs to. α is a constant.

In the definitions of *tf*×*idf*+*CC* and *tf*×*idf*+*OR*, "one-side" metrics (e.g., OR and CC) are employed to embody the feature weight with category discrimination capability. To distinguish positive features from negative features, the positive and negative features are weighted separately, which cannot be accomplished by *tf*×*idf*, *tf*×*CHI* and *tf*×*OR*. For positive features, their *tf*×*idf* weight is added by OR scores or CC scores, and the contributions of positive features to document similarity are increased. Whereas for negative features, their weight in *tf*×*idf+OR* or *tf*×*idf+CC* is equal to their *tf*×*idf* weight.

Additionally, it is obvious that the definitions of *tf*×*idf+CC* and *tf*×*idf+OR* are similar, which indicates that CC or OR can be replaced by other "one-side" feature selection metrics.

3.3 tf×FE

Feature entropy (FE) is a simple metric and contains the category information. FE measures the diversity of the distribution of a feature in different categories. The formula of FE is defined as follows:

$$FE(t_i) = h + \sum_{j=1}^{|C|} p(c_j) \log p(c_j) . \tag{7}$$

where h is a constant and satisfies the following inequation:

$$h > \left| \sum_{j=1}^{|C|} p(c_j) \log p(c_j) \right| . \tag{8}$$

The second part of formula (7) is the negative entropy of feature t_i. $p(c_i)$ is the probability of feature t_i occurred in c_j , and can be estimate by:

$$p(c_j) = \frac{DF(t_i, c_j)}{N} . \tag{9}$$

where $DF(t_i,c_j)$ represents the document frequency of t_i in c_j, and N is the total number of documents.

If a feature t_i has the same document frequency in distinct categories, then the FE of t_i equals to its minimum value, which means that t_i has no discriminability. The distributions of t_i in categories are more different, the FE is higher. The feature weighting function *tf*×*FE* can be defined as:

$$w_{tf\times FE} = tf(t_i, d) \times FE(t_i) . \tag{10}$$

idf is related to the document frequency of features in the whole collection without considering the membership of documents. FE is related to the document frequency in categories. As a result, *tf*×*FE* reflects the category information and *tf*×*idf* doesn't. To our knowledge, FE has never been used in feature weighting.

4 Experiments

In the following experiment, the proposed feature weighting functions are compared with *tf*×*idf*, *tf*×*CHI*, and *tf*×*OR*. The SVM classifier is used in the experiments, and it is provided by LibSVM [15]. Each category is treated as a binary classification problem. All feature weighting approaches are evaluated with both the linear kernel and the RBF kernel. The parameters of kernels are set to the default values in LibSVM.

Data Collection. The evaluating data set used in our experiments is the Reuters-21578 corpus, which is widely used in text classification tasks and has become a benchmark. The version is the ”ModApte” split, which has 90 categories, consisting of 7770 training documents, some of which have multiple category labels, and 3019 test documents. The most frequent ten categories of ninety are used.

After stemming and stop words removing, the unbalanced training set has 24,329 features left, and all of them are used.

Performance Measures. To measure the performance of feature weighting methods, the micro-averaging break-even point (BEP), micro-averaging F_1 and are macro-averaging F_1 [6] used in the experiments.

Result. Table 1 and Table 2 show the results of SVM classifiers with RBF kernel and linear kernel respectively. It list micro-averaging BEP values for each of the most frequent ten categories, and micro-averaging BEP, micro-averaging F_1 and macro-averaging F_1 for all ten categories. We can observe that SVM classifiers with linear kernels have higher accuracy than the classifiers with RBF kernels. For linear kernel, *tf*×*FE* is the best weighting method, and improves *tf*×*idf* with 2.29percent on macro-averaging F_1. *tf*×*FE*, *tf*×*idf*+*CC* and *tf*×*idf* perform better than *tf*×*idf*, *tf*×*OR* and *tf*×*CHI*. For RBF kernel, the performance of *tf*×*FE* is disappointing. However, *tf*×*idf*+*CC* is the best. Both *tf*×*FE* and *tf*×*CHI* are affected by the unbalancedness of Reuters-21578 corpus.

tf×*idf*+*CC* and *tf*×*idf*+*OR* are more effective than *tf*×*OR*, *tf*×*CHI* and *tf*×*idf*. The results confirm that positive features should be distinguished from negative features, and positive features’ contribution should be enlarged. *tf*×*idf*+*CC* is more robust than others on the unbalanced dataset with different kernels.

Table 1. Performances of feature weighting methods with RBF kernel on Reuters-21578

Category	tf×idf	tf×CHI	tf×OR	tf×idf+ OR	tf×idf+ CC	tf×FE
Acq	84.39	42.36	79.53	87.55	**90.82**	76.71
Corn	62.50	56.25	75.89	66.48	**87.16**	50.00
Crude	72.46	50.53	73.57	74.83	**83.66**	57.88
Earn	96.53	56.39	95.49	96.71	**97.79**	94.59
Grain	74.92	55.37	78.42	79.81	**87.53**	54.36
Interest	58.80	53.44	57.16	62.45	**68.14**	51.16
Money-fx	58.68	52.51	56.85	59.33	**67.25**	46.25
Ship	58.19	51.12	59.53	58.19	**72.16**	50.00
Trade	68.08	25.43	70.52	70.16	**77.39**	56.17
Wheat	70.27	54.93	76.63	76.25	**89.06**	54.93
micro-averaging BEP	83.10	52.89	81.97	84.94	**89.24**	75.51
micro-averaging F_1	80.68	13.50	79.70	83.27	**88.88**	69.50
macro-averaging F_1	65.51	11.86	66.86	67.78	**81.33**	35.73

Table 2. Performances of feature weighting methods with linear kernel on Reuters-21578

Category	tf×idf	tf×CHI	tf×OR	tf×idf+ OR	tf×idf+ CC	tf×FE
Acq	95.36	92.46	95.02	**95.50**	95.29	95.30
Corn	86.30	85.71	**91.26**	87.30	88.50	88.14
Crude	84.83	79.18	85.11	86.02	**88.34**	86.64
Earn	97.83	96.01	97.26	97.96	97.54	**97.97**
Grain	87.65	90.13	90.17	90.09	**93.25**	89.67
Interest	71.23	64.55	70.42	**72.65**	71.77	72.31
Money-fx	72.74	70.07	**77.32**	73.81	73.84	76.05
Ship	79.12	70.35	76.57	80.40	**82.54**	81.17
Trade	64.49	58.30	69.74	66.75	66.00	**73.44**
Wheat	85.00	80.86	85.00	86.72	**89.11**	86.72
micro-averaging BEP	90.03	87.02	90.28	90.70	90.91	**91.29**
micro-averaging F_1	90.02	87.00	90.24	90.70	90.90	**91.30**
macro-averaging F_1	82.45	78.75	83.73	83.72	84.59	**84.74**

Discussion. From the definitions of OR and CC, we can observe that positive features selected by OR and CC are exactly the same. Because for a feature t_i and a category c_j, if $AD > CB$, then $OR(t_i,c_j) > 0$ and $CC(t_i,c_j) > 0$. But the scores assigned by OR and CC are different, moreover the rank of features according their OR scores are different from their CC scores. Consequently, *tf*×*idf*+*OR* and *tf*×*idf*+*CC* have distinction on their performances.

5 Conclusion

In this paper, three new feature weighting methods are proposed. *tf*×*idf* is combined with "one-side" feature selection functions in a moderate way and the performance of *tf*×*idf* is improved. Then, FE is proposed, which is concise and effective. The experimental results show that, for the SVM classifier with linear kernels, *tf*×*FE* outperforms other methods. All the three proposed methods are more favorable than *tf*×*CHI*, *tf*×*OR* and *tf*×*idf*.

In the future work, we will investigate the effect of more feature selection metrics on feature weighting methods. And the helpfulness of other metrics to feature weights will be explored, for example, the number of documents in each category, the number of categories that containing a term.

References

1. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: European of Conference on Machine Learning, Chemnitz, pp. 137–142 (1998)
2. Yang, Y., Chute, C.G.: An example-based mapping method for text categorization and retrieval. ACM Transactions on Information Systems 12, 252–277 (1994)

3. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text categorization. In: IJCAI 1999 Workshop on Machine Learning for Information Filtering, Stockholm, pp. 61–67 (1999)
4. Schapier, R.E.: Boostexter: A boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
5. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: International Conference on Machine Learning, pp. 412–520 (1997)
6. Sebastiani, F.: Machine learning in automated text categorization. Computing Surveys 34, 1–47 (2002)
7. Zheng, Z.H., Wu, X.Y., Srihari, R.: Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter 6, 80–89 (2004)
8. Zheng, Z.H., Srihari, R., Srihari, S.: A feature selection framework for text filtering. In: 3rd IEEE International Conference on Data Mining, Melbourne, pp. 705–708 (2003)
9. Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. Studies in Fuzziness and Soft Computing 138, 71–98 (2004)
10. Deng, Z.H., Tang, S.W., Yang, D.Q., Li, L.Y., Xie, K.Q.: A comparative study on feature weight in text categorization. In: Yu, J.X., Lin, X., Lu, H., Zhang, Y. (eds.) APWeb 2004. LNCS, vol. 3007, pp. 588–597. Springer, Heidelberg (2004)
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management 24, 513–523 (1988)
12. Rijsbergen, V.: Information Retrieval. Butterworths, London (1979)
13. Mladenic, D., Grobelnik, M.: Feature selection for classification based on text hierarchy. In: Conference on Automated Learning and Discovery, the Workshop on Learning from Text and the Web, Pittsburg (1998)
14. Ng, W., Goh, H., Low, K.: Feature selection, perceptron learning, and a usability case study for text categorization. ACM SIGIR Forum 31, 67–73 (1997)
15. Chang, C., Lin, C.: LibSVM: a library for support vector machines, `http://www.csie.ntu.edu.tw/cjlin/libsvm`

Algorithms of BBS Opinion Leader Mining Based on Sentiment Analysis

Xiao Yu, Xu Wei, and Xia Lin

Department of Electronics and Information Engineering,
Huazhong University of Science and Technology, Wuhan, China
{xiaoyu,xuwei,xialin}@mail.hust.edu.cn

Abstract. Opinion leaders play a crucial role in online communities, which can guide the direction of public opinion. Most proposed algorithms on opinion leaders mining in internet social network are based on network structure and usually omit the fact that opinion leaders are field-limited and the opinion sentiment orientation analysis is the vital factor of one's authority. We propose a method to find the interest group based on topic content analysis, which combine the advantages of clustering and classification algorithms. Then we use the method of sentiment analysis to define the authority value as the weight of the link between users. On this basis, an algorithm named LeaderRank is proposed to identify the opinion leaders in BBS, and experiments indicate that LeaderRank algorithm can effectively improve the accuracy of leaders mining.

Keywords: social network, Opinion Leader, community discovery, sentiment analysis.

1 Introduction

Users in BBS usually initiate or reply to topics which they interested in. And there are more interactions between users with similar interests than others. BBS users have a natural characteristic of interest clustering, which means people who have similar interests will discuss together. So BBS is actually divided into several independent interest fields. From the above analysis of BBS's characteristics, we construct the network model of BBS. As shown in Figure 1,each circle represents a topic, each square represents an interest, each dashed line in user layer represents a reply relationship and each oval in user layer represents an interest group. Users generate articles because of their interests, and one user can have several interests. In order to build our network in real BBS, we start our study from topic layer. We cluster topics to identify interests, and find interest groups in users.

During a period of time, a small group of users will become core role in a certain interest field. They have many followers and often put forward compelling ideas. These core users are called "opinion leaders". Opinion leaders initiate or participate in popular topics in BBS, and attract a large number of people to participate in their discussions. They strongly impact people around. In general, identifying opinion leaders from BBS can serve many useful purposes, including but not limited to better understand the public opinions of their fields, and it can also help to capture key concerns and potential trends among people. Identifying opinion leaders is very important and meaningful.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 360–369, 2010.

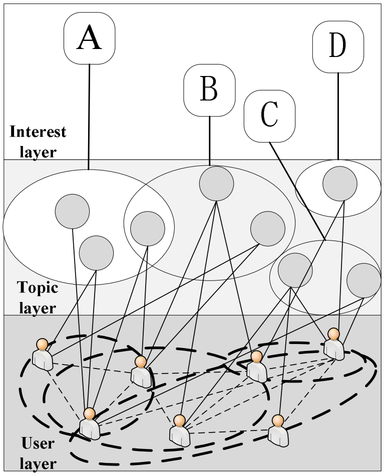


Fig. 1. Structure of social network on BBS

In literature, many measures such as Indegree, PageRank, online time and experience value, have been extensively used to identify leaders from social networks. But they show obvious limitations, for example, ignoring their sphere of influence and semantic information hidden in social exchanges. To solve these problems, we introduce the concept of interest groups, and apply a PageRank-like algorithm named LeaderRank in interest groups for opinion leaders mining. The rest of this paper is organized as follows. In Section 2 we present some related works briefly. Interest groups identifying and sentiment analysis are described in detail in section 3. Section 4 gives a brief introduction to dataset and explains the LeaderRank algorithm and the evaluation metric for algorithms. The experiment results and discussion are presented in Section 5. We conclude this paper in Section 6 with a summary.

2 Related Works

Scholars have conducted widely research on identifying opinion leaders in social networks, and they have proposed a lot of algorithms to identify the opinion leaders and organizations in virtual community. Some important algorithms based on network hyperlink structure analysis, such as PageRank and HITS, have been used as the basis of important theoretical models. These models have been introduced to researches on relationships recognition in the text-based communication network [1]. Jun Zhang et al used Java Forum network as the research object, and they used HITS algorithm and several other algorithms to evaluate users' authority [2]. Hengmin Zhou et al studied opinion mining in opinion network, and proposed an opinion leaders identifying algorithm based on sentimental analysis. But they didn't consider that the impact of opinion leaders would be restricted in interest field [3]. Zhongwu Zhai et al proposed interest-field based algorithms which not only took into account of the reply network's structure but also the users' interest field, and they found that the interest-field based algorithms are sensitive to the high status nodes in the communication network [4].

These studies simplify the social network model to the relationships between repliers and authors. But the following problems are ignored:

- Most of the studies ignored the weight brought in by multiple interactions between users. Their studies were based on an unweighted network, which did not take into account of the weight of edge between network nodes.
- Edge weight cannot be equal to the number of replies author received from one replier, because negative replies will reduce author's authority. Thus, sentiment analysis is needed to calculate whether replies improve the author's authority or not.
- Opinion leaders exist in certain interest field. Some experts are not popular in other field they not interested in. Their authorities are high only in their active area or their interest field.

Therefore, this research will focus on analyzing user-generated topics, and identify interest groups which are composed of users discussing similar topics. Then we use sentiment analysis method to obtain the real authority value of users. At last we propose an opinion leaders identifying algorithm based on community discovery and sentiment analysis.

3 Interest Groups Identifying and Sentiment Analysis

3.1 The Discovery and Identification of Interest Groups

Social network is a typical complex network, and one of its important features is the community structure. A large amount of studies have shown that social network was heterogeneous, and composed of some nodes of the same type. There are more connections between similar nodes than between different nodes. Sub-graphs composed of nodes of the same type and edges between these nodes are communities in network. Community discovery in complex networks originated in the sociological research [5].Wu and Huberman's research [6] and Newman and Girvan's [7] research, made community discovery become an important direction of complex network research in recent years.

Identifying BBS interest groups is implemented by gathering users who post topics about similar interests. User initiates a topic or leaves a reply article to make interactions with other users, so interest groups identification is essentially a text-based classification process. It's impractical to predict how many interest categories exist in massive topics, so automatic text clustering method can effectively merge similar topics to a topic group and decide what the main idea of these gathered topics. Clustering algorithm can be theoretically used to identify interest groups, but in practice it doesn't work well. For example, there are some keywords with high differentiation such as "Maradona" and "River Plate" in an article, human can immediately distinguish this article is about football, but the word like "football" never appeared in this article. It's impossible for pure word-based text clustering algorithm to carry out in-depth intelligent analysis of the semantic context, thus generating to too many topic groups. This is difficult to overcome for a clustering algorithm when dealing with similar issues.

The algorithm based on text classification can effectively avoid this problem. But classification algorithm is a method based on prior knowledge of the field, preconfigured categories and discrimination threshold are needed. So this paper proposes an automatic classification algorithm under the guidance of the clustering algorithm, which can effectively solve problems about interest categories and interest groups identification. Clustering algorithm used in this study refers to 1NN text clustering algorithm and the core ideas consult [8].The core ideas are as following:

- Identify feature items groups of articles and make classification under the guidance of automatically clustering algorithm. Calculate TFIDF vector for each article, and then set threshold to 0.655, at last use 1NN text clustering algorithm to generate topic clusters.
- Artificially amalgamate topic clusters in the same interest field into a single topic group. And adjust feature items group of the new topic group according to the IDF value.

According to topic groups and feature items groups from step 2), use support vector machines classification algorithm to automatically classify massive articles. The ownership threshold S is set to 3.0.

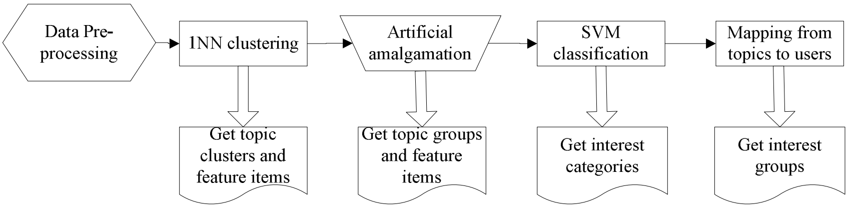


Fig. 2. Process of interest groups identifying

Massive topics can be effectively classified through steps above. This algorithm is efficient and accurate. The only imperfection is the need for manual guidance to amalgamate similar topic groups to interest category. However, after clustering the number of topic groups has been greatly reduced, the workload is small. So in practice, this method is still very effective.

Assemble authors and repliers, who participate in topics of a certain category, we can obtain interest groups.

3.2 Sentiment Analysis of Replies

The emotion of a reply can reveal the replier's positive, negative or neutral attitude to author. Shen Yang constructed an affective words dictionary to help mining the hidden emotion from micro-blog. Test results were cross-checked and reached an accuracy fate of 80.6% [9]. According to the characteristics of BBS, we adjust the algorithm mentioned above.

- Retaining the core structure of the algorithm, we emend affective words dictionary, and introduce custom negative words dictionary, degree words dictionary and exclamation dictionary. On the basis of BBS emotion words analysis, we add 514 affective words, and we set polarity and strength of 1852 words. At last, we normalized words' weights to [-1, 1].

- Reply articles in BBS, different from message in micro-blog, are not limited to 140 words. And paragraphs may exist in a reply. So we enhance the weight value of not only the first and last sentences but also the first and last paragraphs.
- If user B replies to user A for several times, B's final attitude to A should be an average emotion value of those replies.

To verify the effectiveness of the adjusted sentiment analysis algorithm in BBS environment, we crawled date from a famous BBS in China (http://bbs.ccnu.edu.cn). Three days' date, 3128 articles, was chosen to generate test samples. Discard 49 invalid articles like ads and links, the remaining 3079 articles are test samples. Compared the results calculated by adjusted algorithm to artificial results from five students, accuracy of calculation shows below.

Table 1. Sentiment Analysis Accuracy

Day	Emotion	Artificial result	Algorithm result	Accuracy (%)
Day1	Positive	577	436	75.5
	Negative	435	359	82.5
	Neutral	129	82	63.6
Day2	Positive	731	566	77.4
	Negative	334	262	78.4
	Neutral	243	167	68.7
Day3	Positive	429	313	72.9
	Negative	117	89	76.1
	Neutral	84	52	61.9

The result shows that adjusted sentiment analysis algorithm can correctly identify 75.3% positive articles, 79.0% negative articles and 64.7% neutral articles. The negative emotion identification is more accurate, because negative words dictionary contains most of the negative words. Our algorithm will work better with more professional words dictionaries. As neutral emotions identification is not accurate enough, we expand the acceptance boundary of neutral emotion by 0.1, which can get the better result in practice.

Whether user B think user A is a person of authority depends on an integrated impression which user A makes on user B. In this study, W_{AB}, average emotion value of user B's replies to user A, is used to represent A's authority value on B. Calculate authority value between them, and finally we get the authority value matrix.

$$\mathbf{W} = \begin{bmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{n1} & \cdots & W_{nn} \end{bmatrix} \tag{1}$$

$$W_{ij} = \frac{\sum e_{ij}}{t_{ij}} \tag{2}$$

e_{ij} is the emotion value of replies from user j to user i in an article chain, and $\sum e_{ij}$ is the summation of emotion values in all article chains. t_{ij} is the times user j replies to user i.

3.3 Identification of Opinion Leaders

Opinion leaders exist in their own interest area. For example, experts in football may not be experts in military. So the first step to identify opinion leaders is to indentify interest areas. Use the method mentioned above we can get interest groups, and one user can belong to different groups. We calculate authority value matrix in interest group, and use this matrix to find opinion leaders. The flow chart shows below.

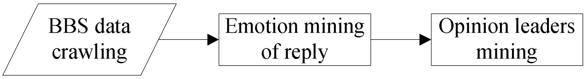


Fig. 2. Process of opinion leaders mining

3.3.1 Data Set Analysis

We crawled articles post between January 2007 and December 2009 from a university BBS to launch our research. We discard replies that authors reply to themselves. We treat those who get replied as authors no matter whether they initiate topics or not.

Table 2. Statistics Of the BBS Network

Statistics subject	value
Number of registerrd users（ *U*）	49902
Number of active users（ *N*）	12779
Number of boards（ *G*）	120
Number of articles（ *W*）	906633
Mean degree（ *K*）	52.71
Average cluster coefficient（ *C*）	0.94
Average path length（ *L*）	3.07

Table 2 shows some social network characteristics of the sample data. The average degree reaches 52.71, it means that this BBS is a popular forum.

We chose data from the biggest board to analyze. There are 19687 articles and 2215 users in this board. We discard the isolate authors who have no reply. 128 topic groups are generated after elementary clustering. We use artificial method to amalgamate these topic groups to 9 categories. The quantity of each category shows below:

Table 3. Result Of Topics Classification

Category name	Number of topic	Proportion(%)
Teacher	11844	49.66
Tuition fee	3413	14.3
education	1841	6.88
university	1326	5.53
study	347	1.43
enrollment	288	1.17
exam	264	1.09
career	184	0.76
other	176	0.71

Topics about similar interests make up a category. We collect authors and repliers related to these topics and then form interest groups. Take the category named "teachers" for example, and we draw an interpersonal relationship graph of this interest group. It's obvious that this network is heterogeneous. There are more connections between core nodes, and a few connections between the nodes around. Small groups are clear in the graph.

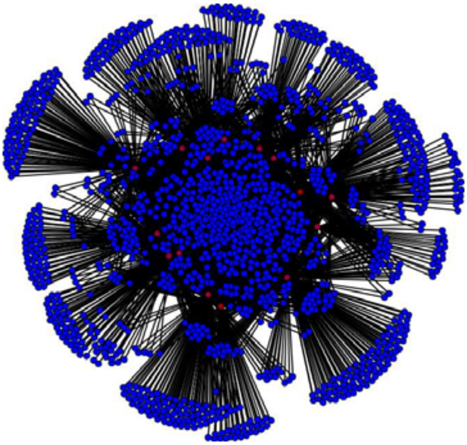


Fig. 4. Social network structure in an interest group

3.3.2 Leaderrank Algorithm Based on Sentiment Analysis

The well known PageRank algorithm, proposed by Page et al, for ranking web pages, has been widely used in social network for opinion leaders mining. We apply a PageRank-like algorithm to identify opinion leaders. This measure, called "LeaderRank", not only takes into account of the interest field of opinion leaders but also replier's attitude to author. LeaderRank shows as the following formula:

$$LR(u) = (1 - d) + d * \sum_{v \in B_u} \frac{LR(v) * w_{uv}}{C(v)} \tag{3}$$

$$C(V) = \sum_{k \in T_v} |w_{vk}| \tag{4}$$

LR(u) is node u's leader rank score; B_u is the set of nodes linked to u; T_v is the set of nodes linked by *v*; w_{uv} is the authority value from *v* to *u*; *C(V)* is the sum of *v*'s out-link-weights' absolute values. *d* is the damping coefficient. The value of *d* is set to 0.85. The LeaderRank score of each user is initiated to 0.1, and then is iteratively updated until the scores converge.

In order to test LeaderRank algorithm, we compare the result of LeaderRank with results of other five traditional algorithms. Some studies show that traditional algorithms do well in opinion leaders mining [2,4]. These traditional algorithms are:

- **Indegree:** use indegree to rank users.
- **Global PageRank:** run PageRank algorithm in the whole board.
- **Interest-based PageRank:** run PageRank algorithm in interest field.
- **Online time:** rank users by their online time.
- **Experience value:** use experience value BBS system automatically assigns to rank users.

3.3.3 Evaluation Metrics

There is no explicit opinion leadership rank system. According to the definition of opinion leader, opinion leaders interact with a large amount of users, and the frequency of interactions is very high. So we propose a metric named "core radios" as following:

$$\mathrm{CR(i)} = \frac{\sum_{j=1}^{N} a_{ij} w_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{i} a_{ij} w_{ij}} \tag{5}$$

W_{ij} is the weight of the edge between user i and user j, and its value is equal to user i's authority value on user j. When user j replies to user i a_{ij} is set to 1, and when user j never replies to user i a_{ij} is set to zero. Core ratios are calculated in each interest field at first, and the average of these ratios is the final core ratios of each algorithm.

4 Results and Analysis

We compare the core ratio of opinion leaders identified by each algorithm. The average core ratios of top K users are shown in Figure 5.

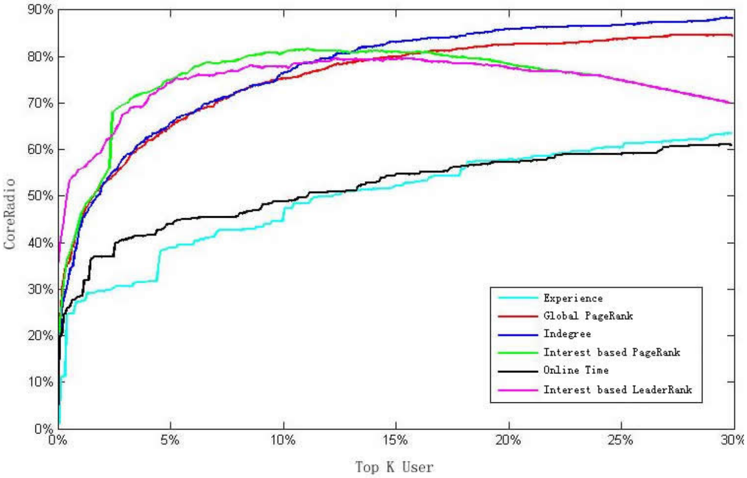


Fig. 5. Core ratios of six algorithms

We can find from figure 5 that four kinds of algorithms have good correlation. They are LeaderRank algorithm, interest-based PageRank algorithm, PageRank algorithm and Indegree. And these four algorithms do much better than experience value and on-line time. This shows structural characteristics of the network can by very useful in opinion leaders mining. The first four methods have a common feature: their curves increase before the point 10%, and become flat after this point. This phenomenon shows 10% of users cover 80% of the interactions, and it means less than 10% of users can be opinion leaders. We find that two kinds of interest-based algorithms are better than other algorithms, indicating that identify interest fields can help finding opinion leaders. This is consistent with citation [4]. Experience value cannot serve as an accurate opinion leader mining method, because it only takes into account of the amount of articles users post but not the relationship between users. LeaderRank algorithm has outstanding performance at the point of 5%, that is to say LearderRank can find prominent opinion leaders more quickly than other algorithms.

5 Conclusion

This research conducts an in-depth analysis of the form of network community on BBS, and proposes a network model based on interest field. We also propose approach of community discovery by classification algorithm guided by results of text clustering to classify topics and identify interest groups. On this basis, we use sentiment analysis algorithms to calculate the authority value of one user on another. Then we take authority value into account and propose a LeaderRank algorithm to identify opinion leaders in interest field. Our experimental results suggest that interest cluster and sentiment analysis do have strong impact on social network analysis.. This discovery is important and it is helpful to better understand the formation of public opinion.

Acknowledgment

This work is supported by Technology Support Plan of the National ’Eleventh Five-Year-Plan’ of China under Grant No. 2006BAK11B00. We also wish to thank Cheng Wenqing, Xu Jing for their help, and anonymous reviewer for feedback and suggestions.

References

1. Matsumura, N., Ohsawa, Y., Ishizuka, M.: Influence Diffusion Model in Text-Based Communication. In: WWW 2002 (2002)
2. Zhang, J., Ackerman, M., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: J. WWW 2007 (2007)
3. Zhou, H., Zeng, D., Zhang, C.: Finding Leaders from Opinion Networks, pp. 266–268 ISI (2009)
4. Zhai, Z.W., Hua, X.: Identifying opinion leaders in BBS. J. IEEE Proceedings of Web Intelligence and Intelligent Agent Technology (2008)
5. Scott, J.: Social Network Analysis:A Handbook. Sage Publications, London (2000)
6. Wu, F., Huberman, B.A.: Finding communities in linear time: A Physics approach. J. Euro, Phys. J. B 38, 331–338 (2003)
7. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. J. Physical Review E (2004)
8. Cover, T.M., Hart, P.E.: Nearest Neighbor Pattern Classification. IEEE Trans, Information Theory, 21–27 (1967)
9. Yang, S., Li, S., Zhen, l.: Emotion mining research on micro-bolg. SWS, 71–75 (2009)

Entity Relationship Extraction Based on Potential Relationship Pattern

Chen Chen[1], HuiLin Liu[1], GuoRen Wang[1], LinLin Ding[1], and LiLi Yu[2]

[1] Dept of Computer Science and Technology, Northeastern University,
Shenyang 110004, China
neuchenchen@163.com, liuhuilin@ise.neu.edu.cn
[2] Mobile Internet Business Unit, Neusoft Corporation
Shenyang 110179, China
yu.ll@neusoft.com

Abstract. The keep rising of web information ensures the development of entity focused information retrieval system. However, the problem of mining the relationships effectively between entities has not been well resolved. For the entity relationship extraction (RE) problem, this paper firstly establishes the basic pattern trees which can present the overall relation structures and then designs a similarity function according to which we can judge which pattern the sentence containing two entities belongs to. Knowing the matched pattern, we can discovery the relationship easily. By a large number of experiments on real data, the proposed methods are proved running accurately and efficiently.

Keywords: entity; relationship extraction; pattern; similarity function.

1 Introduction

Along with the popularization of information technology and rising of web recourses, the change happens for both the server interface of a retrieval system and the information requirement of users. For example, at present the entity focused retrieval becomes the hot issues. It is a new retrieval mode, there are many problems need to be resolved, such as entity recognition, attribute mining, entity ranking, co-reference resolution and so on. These problems are all considered from the point of single entity. In fact, entity relationship plays more important role in many circumstances. For example, for a Question & Answer system, a complete relationship knowledge database will help to answer the question "Who is the Chairman of the Olympic Games?". Because of its importance, researchers have proposed many effective methods. However, most of their researches are based on several fixed relationships. For example, given the relationship keywords "locate in", their work can judge whether two entities in a sentence contain such relationship. Under large web scale, it is unpractical to define all the relationships previously. Without any extra knowledge needed, the paper will extract any potential relationships between entities directly.

In the paper, all the considered entities exist in a common sentence and all the sentences mentioned following have labeled two given entities if no special instructions.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 370–377, 2010.

By the analysis of a large number of text , there are several important patterns presenting the syntactic structures and these patterns can be used for predicting the relationship keywords. So to resolve the open RE problem, we first gave a brief introduce for the work of Michele Banko[1] who summarized eight basic syntactic structures for depicting relationships and then define the concept of pattern tree. If we can ascertain which pattern a sentence belongs to, then we'll get the relation words easily.So based on the pattern tree, we propose a similarity function which considers the content similarity and location similarity together. Use this function, the pattern with maximum value will be the answer.

The structure of the rest paper is as follows. In section 2 we describe some related work. In section 3, we'll illustrate the eight basic relationship patterns and their tree presentations. The detail computations between patterns and given sentences will be presented in section 4. We describe the experimental results on section 5 and summarize our conclusions and suggest future directions in section 6.

2 Related Work

RE is a subtask in the domain of entity retrieval. There are many related works. The most outperform methods are HMM [2] and maximum entropy-based [3] method.

Based on entity recognition, relationship extraction was first proposed in MUC[4], and so far there are many achievements. Among these, the most popular are the feature-based method and kernel-based method.

The feature-based method utilizes the type feature, syntactic feature as well as semantic feature to train the relationship extraction model. Focused on the problem of relation extraction of Chinese entities, reference [5] proposed nine position structure features and trained a model under the dataset of ACE2005[6]. Kambhatla[7] presented the features by the concept of maximum entropy. Based on his work, Zhou[8] extended the method, involving more features by introducing lineal kernel function and SVM.

Relative to feature-based method, the kernel-based method can utilize the structure information wonderfully. In [9], authors used parse tree to present sentences and computed the similarity of two trees recursively. Culotta[10] computed the similarity of two sentences utilizing dependency tree. Reference [11] defined the concept of shortest path and got better results. Reference [12] pointed out the shortcomings of shortest path and proposed the context-sensitive structured parse tree.

All the work has obtained many achievements, but as analyzed in section 1, those methods always concern several certain relationships and can't be used in web environment directly. So our work is significant.

3 Relationship Pattern Tree

The natural language has powerful presentation ability. The same meaning can be depicted by different sentence types and the same sentence type can express different messages. For the RE problem, we assume that for the two sentences with same syntactic structures, the relation keywords will also exist in the same position. Based on the hypothesis, for a pattern set, if we can assign a fixed pattern for each sentence, we'll find the relation words easier.

In [1], Michele Banko and Oren Etzioni have analyzed amount of relationship instances and summarized eight main relationship patterns. The statistical results of these patterns are described as table 1.

Table 1. The basic patterns

Frequency (%)	Category description	Simple instances	
37.8	Verb	E1 Verb E2	(X established Y)
22.8	Noun + Prep	E1 NP Prep E2	(X settlement with Y)
16.0	Verb + Prep	E1 Verb Prep E2	(X moved to Y)
9.4	Infinitive	E1 to Verb E2	(X plans to acquire Y)
5.2	Modifier	E1 Verb E2 Noun	(X is Y winner)
1.8	Coordinate(n)	E1 (and|,|-|,) E2 NP	(X, Y deal)
1.0	Coordinate(v)	E1 (and|,|-|,) E2 Verb	(X, Y merge)
0.8	Appositive	E1 NP (:|,)? E2	(X hometown: Y)

The first column of table 1 shows the number of occurrences of each pattern. Column two is a short summary for the syntactic structure of each pattern. The last column gives simple instances. Add the value in first column, we get 94.8%. Therefore, table 1 covers the majority situations of presenting relationships. To utilize the resources effectively supplied by table 1, we convert them into the tree structures. By the open source project Opennlp[13], we can get corresponding parse tree structures as figure 1 (Only the first 3 patterns are illustrated).

In figure 1, each tree stands for a specific syntactic structure and we call them pattern trees. In the pattern tree, leaf nodes stand for the real contents and other nodes stand for the syntactic structures. Such as NP is none phase, IN is prepositional phrase, VBD is transitive verb and so on. The X and Y in leaf nodes are two labeled entities.

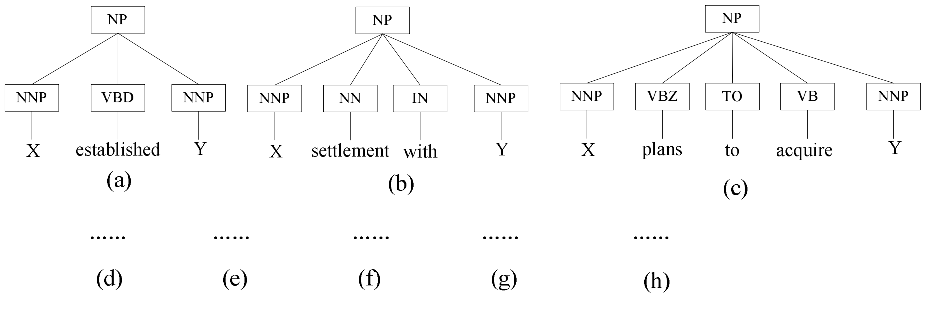


Fig. 1. The eigh basic pattern trees

Analyze these pattern trees, we can find that there are fixed grammar units presenting the relationships, such as the verb “established” in figure 1(a), and the prepositional phrase “settlement with” in figure 1(b). Therefore, the RE can be considered as a classification problem. If we can assign an accuracy category to given sentence, then it will be easy to assure the relation words.

4 Relationship Pattern Matching

According to the hypothesis, any sentence will attach to one of the patterns. This section will illustrate the process of relationship pattern matching. In the paper, we use conditional probability to compute the similarity between sentences and patterns. For a given sentence, the probability of the pattern it belongs to can be depicted as formula 1.

$$C = \arg\max_{i} p(c_i \mid s), \qquad i \in [1-8] \tag{1}$$

In formula 1, i stands for the id of pattern, $i \in [1-8]$; s stands for the sentence need to be recognized. c_i stands for the ith pattern. For each sentence, we compute eight values for $i \in [1-8]$, the pattern with maximum value will be the answer.

Shown as figure 1, the pattern trees stand for the basic syntactic structure of sentence, which have simple layer structures. However, a real sentence will be more complex, which can be viewed as extending of basic patterns. In the paper we will use content and location together to determine the similarity.

In a parse tree, we think that the nodes which connect the leaves directly always have the obvious structure information. Collect this kind of nodes together, we get the pattern vector, short for SV. For distinction, we use SV_c stands for the pattern vector of pattern and SV_s for sentence. For example, in figure 2 we get the pattern vectors as follow.

SV_c(a)={NNP,NN,IN,NNP}, // the pattern vector for figure2 (a);
SV_c(b)={NNP,CC, NNP,NN}, // the pattern vector for figure2 (b);
SV_s(c)={NNP,CC,NNP,VBP,DT,NNP,NN,IN,NNP},
// the pattern vector for figure2(c);

We'll use these vectors to compute the similarity.

The goal of this paper is to extract the relation words of two given entities, so we can ignore the useless grammar units in a sentence. Ideal condition, a sentence contains only entities and relation words will be easiest to compute. So, if the useless modified unit can be filtered when computation, the proposed method will work more efficient.

Observe the eight patterns in figure 1 and table 1, there are two types of position relationships for entity X and Y. One kind can be presented as X-relation-Y, namely the relation words locate between the two entities, and we call this as embedded mode. Another type can be presented as X-Y-relation, namely the relation words locate behind of the Y entity, we call this as extend mode. For example the a, b , c, d, h in figure 1 belong to the first case and e, f ,g will belong to another.

According to the above definitions, when we compute the similarities between sentences and patterns with embedded mode, only the minimum sub-tree containing two entities are considered and for extend mode we only consider the minimum sub-tree containing the father nodes of two entities and the brother nodes of father nodes. For the instance in figure 2, when computing with pattern (a), only the structures surrounded by dotted lines are considered. So, the pattern vector of sentence becomes:

SV_s(c)={ DT,NNP,NN,IN,NNP};

When computing with pattern (b), the operate vector keeps unchanged.

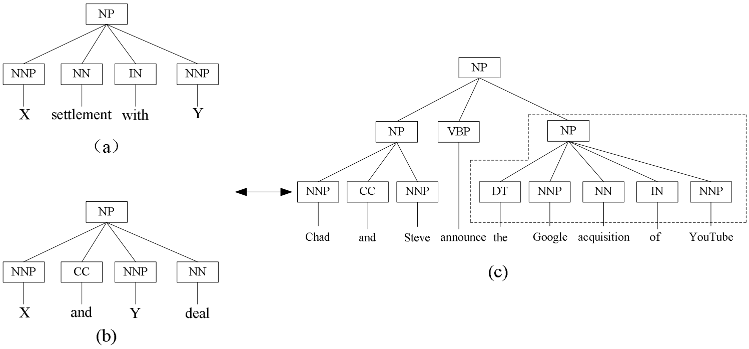


Fig. 2. The diagram of pattern matching

In the paper, the similarity will be considered from two points.

We first define the content similarity as below.

$$CS(\mathrm{SV}_c,\mathrm{SV}_s)=\begin{cases}0; & \text{if } \mathrm{SV}_c \not\subset \mathrm{SV}_s \\ \dfrac{|\mathrm{SV}_c|}{|\mathrm{SV}_s|}; & \text{o.w.}\end{cases} \tag{2}$$

In formula 2, the absolute sign stands for the length of the vector.

Select the elements which also exist in SV_c from SV_s, we construct new vector SV_s' which has the same element sequence as SV_s, then the position similarity can be defined as,

$$PS(\mathrm{SV}_c,\mathrm{SV}_s')=\begin{cases}0; & \text{if } \mathrm{SV}_c \neq \mathrm{SV}_s \\ \left(\sum_{i=1}^{count}\dfrac{|\,substring_i\,|}{|\mathrm{SV}_s|}/\,count\right)\ast\left(\dfrac{1}{\sum|\,p_s(substring_i)-p_c(substring_i)\,|}\right) & ;\ \text{o.w}\end{cases} \tag{3}$$

In formula 3, *substring* stands for the element sequences which have same orders in the two vectors. For example, for the two vectors, V_1={NNP, NNP, VBP},V_2={NNP, VBP,NNP},NNP, VBP will be one *substring*, and |*substring*| is the length of *substring, count* stands for the number of *substring, $p_s(substring_i)$* stands for the location of the *substring* in V_s and $p_c(substring_i)$ stands for the location of the *substring* in SV_c. Such as the location of sequence "NNP, VBP"in V_1 is 2, and is 1 in V_2.

Therefore, Σ |$p_s(substring_i)$-$p_c(substring_i)$|depicts the position difference of *substring* in two vectors.

Take figure 2 as an example. Figure 2(a) and figure 2(b) are two basic patterns of figure 1 and figure 2(c) is the parse tree of "Chad and Steve announce the Google acquisition of YouTube".In the sentence, "Google" and "YouTube" are the two given entities.

According to the definition above, the similarity between figure 2(a) and figure 2(c) is computed as formula (4) and formula (5).

By the results, we can find that pattern (a) gets bigger probability.

$$P(s|c_i)=CS(\mathrm{SV}_c,\mathrm{SV}_s)*PS(\mathrm{SV}_c,\mathrm{SV}_s{}')$$
$$=\frac{|\mathrm{SV}_c|}{|\mathrm{SV}_s|}*(\sum_{i=1}^{count}\frac{|substring_i|}{|\mathrm{SV}_s|}\Big/count)*\frac{1}{\sum|(p_s(substring_i)-p_c(substring_i))|}=\frac{4}{5}*\frac{(\frac{4}{4})}{1}*\frac{1}{1}=0.8 \tag{4}$$

The similarity between figure 2(b) and figure 2(c) is computed as

$$P(s|c_i)=CS(\mathrm{SV}_c,\mathrm{SV}_s)*PS(\mathrm{SV}_c,\mathrm{SV}_s{}')$$
$$=\frac{|\mathrm{SV}_c|}{|\mathrm{SV}_s|}*(\sum_{i=1}^{count}\frac{|substring_i|}{|\mathrm{SV}_s|}\Big/count)*\frac{1}{\sum|(p_s(substring_i)-p_c(substring_i))|}$$
$$=\frac{4}{8}*\frac{(\frac{2}{4}+\frac{1}{4}+\frac{1}{4})}{3}*\frac{1}{|(1-2)|+|(2-3)|+|3-1|}=0.094 \tag{5}$$

5 Relationship Pattern Matching

5.1 Dataset

To prove the availability of proposed method, we implement extend experiments on real dataset. This paper adopts the same dataset of [14]. Everyone can free download it from the address shown in [15]. The dataset contains 500 sentences and each sentence has labeled two entities. For example, <p1> Google </p1> assimilates <p2> YouTube </p2>. In the sentence, the words surrounded by labels p1 and p2 are two entities.

5.2 Experimental Results

To make the structure of the sentence clear, we first convert sentences into the form of parse trees. Then according to the probability formula in section 4, we compute 8 probability values for each sentence and the pattern with maximum value will be the answer. At last according to the definition of pattern, we select specific grammar unit as relation words.

In fact, it is unnecessary to compute eight values for each sentence. The minimum requirement for matching between a sentence and pattern is that $\mathrm{SV_c} \subseteq \mathrm{SV_s}$, namely $\mathrm{SV_s}$ must contain all the elements in $\mathrm{SV_c}$. This restricted condition will filter a large amount of useless computations.

By the probability value, we assign category for each sentence and then find out the relation words. For the extraction results, we judge it's accuracy by artificial way. At last, we get the statistical result as table 2.

From table 2, we can know that the accuracy of proposed method achieves 84.4% (the number of sentences which are extracted out right relation words/500). So under the majority conditions, our method can extract relation words well. Also, there are some phenomenon needs to be noticed.

First, there are only 478 sentences in table 2, 22 sentences don't match with any given pattern. This condition is caused by the character of sentences themselves. Generally speaking, these kinds of sentences have particular syntactic structures and then can't be presented by and given pattern.

Table 2. Relationship extraction results

Id	Pattern mode	Number of sentences	Accuracy	Average accuracy
A	embedded	195	84.5%	87.46%
B	embedded	105	86.6%	
C	embedded	80	82.8%	
D	embedded	42	83.4%	
H	embedded	3	100%	
E	extend	12	58.3%	58.7%
F	extend	36	77.8%	
G	extend	5	40%	
Overrall accuracy =		84.4%		

Second, for the sentence with extend pattern mode; the accuracy is smaller than the sentence with embedded mode. This is because that the syntactic structure of the latter is more complex. For example, in the SV_s vector, there are more useless grammar units which restrain the final extraction accuracy.

Next, we compare our approach to two baseline methods. The majority of existing methods are based on several given relationships. So we first select 3 relationships with most frequency between X and Y in the dataset as below, (X-entity, acquire/ acquisition of , Y-entity), (X-entity, born in, Y-entity) and (X-entity, hometown, Y-entity).

Figure 3 showed the comparison result between baseline methods and our open relation extraction method.

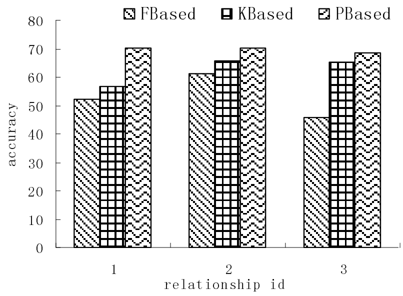


Fig. 3. Comparison result of accuracy

In figure 3, “FBased” stands for the feature-based method, “KBased” stands for the kernel-based method and “PBased” stands for the proposed pattern-based method. Readers can obtain the implementation details of baseline from the references listed in the end of the paper. By the results we can find that our method has better recognition performance.

At last , our method don’t need any extra work before running, so the time consuming and computational complexity are relatively low.

6 Conclusions

RE is the key problem in the domain of entity retrieval and so far has not been resolved well. By the given eight basic relationship patterns, we combine the content similarity and position similarity together and assign the category label for each sentence by seeking the maximum probability value. The experimental result proves the effective of proposed method.

Based on the achievements above, we'll continue our research on the two points as follows.

1. Our method is based on English, compared with English some languages have more complex syntactic structure, such as Chinese, Japanese and so on. We'll continue our work on these languages.
2. Try to construct an entity graph whose vertexes are entities and the edges are relations we have extracted from other resources. The entity graph will provide help for relation retrieval.

Acknowledgment

The work in this paper was supported by the Central Collegiate Basic Scientific Research Bursary Grant N090604011.

References

1. Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. Proceedings of ACL, 28–36 (2008)
2. Lawrence, E., Rabiner, A.: Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
3. Berger, A.L., Pietra, S.A.D., Pietra, V.J.D.: A Maximum Entropy Approach to Natural Language Processing. Computational Linguistic 22(1), 39–71 (1996)
4. http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html
5. Li, W., Zhang, P., Wei, F., Hou, Y., Lu, Q.: A novel feature-based approach Chinese entity relation extraction. In: Human Language Technology Conference (2008)
6. http://projects.ldc.upenn.edu/ace/
7. Kambhatla, N.: Combining lexical, syntactic and semantic features whit Maximun Entropy models for extracting relations. In: ACL 2004, Barcelona, Spain, pp. 178–181 (2004)
8. Zhou, G., Su, J., Zhang, J., Zhang, M.: Exploring Various Knowledge in Relation Exctraction. Proceeding of ACL, 427–434 (2005)
9. Zelenko, D., Aone, C., Richardella, A.: Kernel Methods for Relation Extraction. Journal of Machine Learning Research, 1083–1106 (2003)
10. Culotta, A., Sorensen, J.: Dependency Tree Kernels for Relation Extraction. Proceedings of ACL, 423–429 (2004)
11. Bunescu, R., Mooney, R.: A shortest Path Dependency Tree Kernel for Relation Extraction. In: Proceedings of HLT/EMNLP, pp. 724–731
12. Zhou, G., Zhang, M., Jiq, D.: Tree Kernel-based Relation Extraction with Context-Sensitive Structured Pare Tree Information. In: Proceedings of EMNLP, pp. 728–736 (2009)
13. http://opennlp.sourceforge.net/
14. Sekine, S.: On-demand information extraction. In: Proc.of COLING (2006)
15. http://www.cs.washington.edu/research/knowitall/hlt-naacl08-data.txt

On Arnoldi Method Accelerating PageRank Computations

Guo-Jian Yin and Jun-Feng Yin

Department of Mathematics, Tongji University,
1239 Siping Road, Shanghai 200092, P.R. China
guojian6223@163.com,yinjf@tongji.edu.cn

Abstract. PageRank is a very important ranking algorithm in web information retrieval or search engine. We present Power method with Arnoldi acceleration for the computation of Pagerank vector, which can take the advantage of both Power method and Arnoldi process. The description and implementation of the new algorithm are discussed in detail. Numerical results illustrate that our new method is efficient and faster than the existing counterparts.

Keywords: Information retrieval, Pagerank, Power method, Arnoldi process.

1 Introduction

With the development of internet and its technology, information retrieval and search engine on the web become the most important internet tools, and cause a number of interest from the researchers over the world. Recently, 85% of web users use various internet search engines to find the information from the web. Google is one of the most popular and successful ones.

How to give a order of the webpages according to the importance? As reported in Google's homepage, the kernel technology of Google is its ranking algorithm, called PageRank. It was proposed by S. Brin and L. Page in 1998 [2], and was widely studied in [6,7,8].

In PageRank model, a matrix P is defined from the hyperlink structure of webpages, and then the eigenvector for the dominant eigenvalue of matrix

$$A = \alpha P + (1 - \alpha)E$$

is computed where α is named the dampling factor and E is a rank-one matrix, see [8] for more details. The matrix P can also be treated as a stochastic matrix of a Markov chain. Though P could be reducible. A is irreducible nonnegative matrix with the same largest eigenvalue as P, and the Pagerank vector whose elements sum to 1 is due to the eigenvector associated with the largest eigenvalue of A.

Since the matrix P usually is extremely large (over 8 billion) [3], direct decomposition techniques such as LU and QR decomposition cannot be considered. Iterative methods based on matrix-vector product have been widely studied for the computation of PageRank.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 378–385, 2010.

Power method and its variants attract much attention for computing the PageRank problem. For instance, the quadratic extrapolation method [7], the adaptive method [6], the block method [8]. A number of iterative methods based on Arnoldi process were proposed for computing Pagerank, e.g., Arnoldi-type method [4] and power-Arnoldi method [10]. For a survey, we refer the reader to [3,5].

It is noted that Arnoldi-type method has many advantages. First, the orthogonalization of Arnoldi process achieves effective separation of eigenvectors. In addition, since the largest Ritz value of Arnoldi process could be complex, but if we set the shift to 1, there is no risk of complex arithmetic, and the cost of the algorithm can be greatly saved. Finally, smallest singular value converges smoothly to zero more smoothly than largest Ritz value converges to 1.

Taking advantage of the merit of the Arnoldi-type algorithm, we propose to use Arnoldi-type method to accelerate the Power method for the PageRank computation. The description and implementation of the new algorithm are discussed in detail. Numerical experiments show that the new algorithm is efficient and faster than the existing methods.

The paper is organized as follows. After briefly reviewing Power method for PageRank, we develop Arnoldi accelerated Power method and analyze its properties in details in section 2. Numerical results and comparisons are reported in section 3 and finally in section 4, some brief concluding remarks are given.

2 Arnoldi Accelerated Power Method

In this section, we firstly briefly review the Power method and Arnoldi-type method, and then introduce the Power method accelerated by Arnoldi-type algorithm. The properties and implementations of the new method are studied in details.

The Power method is one of the important iterative methods for finding the dominant eigenvalue and its corresponding eigenvector. The Power method for the computation of PageRank vector is described as follows.

Method 2.01 The Power method

```
Choose $v_0$;
For $k = 1, 2, \ldots$ until convergence, Do:
    Compute $v_{k-1} = v_{k-1}/\|v_{k-1}\|$;
    Compute $v_k = Av_{k-1}$;
EndDo
```

Since the largest eigenvalue of Google matrix is known to be 1, we can computing the residual r by $r_k = v_k - v_{k-1}$ without additional computation for the approximated eigenvalue.

Power method is quite efficient and easy to be implemented. However, matrix-vector multiplication should be computed at every step of the Power method, which is quite expensive because the size of the matrix is usually huge.

A number of acceleration techniques were proposed in past decade, which can be classified into two groups as follows: one is to reduce the total iterative steps, for instance, the quadratic extrapolation method [7], and another is to reduce the computational cost in every step, e.g., adaptive method [6] and block structure method [8]. Unfortunately, these goals usually contradict each other.

However, the convergence performance depends on the gap of the largest eigenvalue and the second largest eigenvalue. When the second eigenvalue of matrix A is close to 1, e.g., when α is close to 1 [9], Power method and its variants still have the difficulty of slow convergence.

Then, some researchers focused their attention on the iterative methods based on Arnoldi process [1]. Given a general matrix $A \in \mathbb{R}^{n\times n}$ and an initial vector $q_0 \in \mathbb{R}^n$, the Arnoldi process gives an orthonormal basis $\{q_1, q_2, \ldots, q_m\}$ of Krylov subspace

$$\mathcal{K}_m(A, v) = \{q_0, Aq_0, \ldots, A^{m-1}q_0\}.$$

Denote $Q_m = [q_1, q_2, \ldots, q_m] \in \mathbb{R}^{n\times m}$, it follows that

$$AQ_m = Q_m H_m + h_{m+1,m} q_{m+1} e_m^T \tag{1}$$

and

$$Q_m^T A Q_m = H_m$$

where $H_m = \{h_{i,j}\} \in \mathbb{R}^{m\times m}$ is an upper Hessenberg matrix. It was suggested that the Arnoldi method can be interpreted as an accelerated Power method [11].

By taking the fact that the largest eigenvalue of A is 1, Golub and Greif [4] proposed an Arnoldi-type method as follows.

Method 2.02 The Arnoldi-type method

```
Choose $q_1$ with $\|q_1\|_2 = 1$
For $i = 1, 2, \ldots, k$ Do
    For $j = 1, 2, \ldots, m$ Do
        Compute $w = Aq_j$
        For $k = 1, 2, \ldots, j$ Do
            Compute $h_{kj} = q_k^T w$
            Compute $w = w - h_{kj} q_k$
        EndDo
        Compute $h_{j+1,j} = \|w\|_2$
        If $h_{j+1,j} \neq 0$
            Set $q_{j+1} = w/h_{j+1,j}$
        EndIf
    EndDo
    Compute SVD: $H_{m+1,m} - [I; 0] = U \Sigma V^T$
    Compute $q_1 = Q_m v_m$
EndDo
```

We should remark that instead of computing the eigenvalues of H_m, a singular value decomposition of $H_m - [I;0]$ is computed in Method 2.02, where all the singular values and singular vectors are real. Note that the smallest singular value of the shifted Hessenberg matrix is not equal to 0; rather, it converges to it throughout the iteration.

Denote σ_m be smallest singular value of $H_{m+1,m} - [I;0]$, then v_m in line 15 is the corresponding right singular vector, and $Q_m v_m$ is the approximated Pagerank vector. Since

$$\begin{aligned} Aq - q &= AQ_m v_m - Q_m v_m \\ &= Q_{m+1} H_{m+1,m} v_m - Q_m v_m \\ &= Q_{m+1} \left[H_{m+1,m} - \begin{pmatrix} I_m \\ 0 \end{pmatrix} \right] v_m \\ &= \sigma_m Q_{m+1} u_m, \end{aligned}$$

it follows that

$$\|Aq - q\|_2 = \sigma_m.$$

Thus, the smallest singular value σ_m can be used to judge the convergence of Arnoldi-type method as stopping rule. For more details and advantages of Arnoldi-type method, we refer the reader to [4].

Our motivation is to take advantage of the merit of either Power method and Arnoldi-type method and seek a balance between two methods. It leads to a powerful approach, named Arnoldi-type accelerated Power method as follow.

Method 2.03 The Arnoldi-type accelerated Power method

```
1.  Choose $v_0$;
2.  For $l = 1, 2, \ldots$ Do:
3.      For $k = 1, 2, \ldots p$ Do:
4.          Compute $v_{k-1} = v_{k-1}/\|v_{k-1}\|_2$;
5.          Compute $v_k = Av_{k-1}$;
6.      EndDo
7.      Set $q_0 = v_p/\|v_p\|_2$.
8.      For $j = 1, 2, \ldots, m$ Do:
9.          Compute $w = Aq_j$;
10.         For $i = 1, 2, \ldots, j$ Do:
11.             Compute $h_{ij} = (w, q_i)_2$;
12.             Compute $w = w - h_{ij} q_i$;
13.         EndDo
14.         Compute $h_{j+1,j} = \|w\|_2$;
15.         If $h_{j+1,j} \neq 0$
16.             Set $q_{j+1} = w/h_{j+1,j}$;
17.         EndIf
18.         Compute SVD: $\overline{H}_m - [I;0]^T = U\Sigma S^T$
19.         Compute  $q = Q_m s_m$
20.     EndDo
```

```
21.     Set $v_0 = q$.
22. EndDo
```

Given an initial vector v_0, the main mechanism of the new approach can be described briefly as follows: we first run p steps of Power method, and then use the approximation Pagerank vector as the initial vector for Arnoldi process; after m steps of Arnoldi-type method, we take the approximation Pagerank vector as the initial vector for Power method.

It is noted that in method 2.03, p steps power iteration in lines 3-6 is performed before m steps Arnoldi-type method in lines 8-20 and exchange the information of approximation Pagerank vector in line 7 and line 21.

The resulting algorithm can be viewed as accelerating Power method with Arnoldi-type method. The following theorem also suggests that the new technique also can be interpreted as Arnoldi-type method preconditioned with Power method.

Theorem 1. [12] *Let w_1 be the left eigenvector of A associated with λ_1 and assume that $\cos\theta(w_1, v_0) \neq 0$. Let q_1 be the orthogonal projector onto the right eigenspace associated with λ_1, i.e.,: $P_1 = q_1 q_1^H$, and B_1 be the linear operator $B_1 = (I - P_1)A(I - P_1)$. Define*

$$\epsilon_m = \min_{\substack{p \in \mathbb{P}_{m-1} \\ p(\lambda_1)=1}} \|p(B_1)(I - P_1)v_0\|_2$$

Then, we have

$$\|(I - \mathcal{P}_m)q_1\|_2 \leq \frac{\epsilon_m}{|\cos\theta(w_1, v_0)|}$$

where v_0 is the generating vector of subspace $\mathcal{K}_m(A, v_0)$ that Arnoldi procedure projects onto, and $\mathcal{P}_m$ denotes the orthogonal projector.

The theorem indicates that the performance of Arnoldi-type method for computing PageRank vector is closely related to the initial vector. Running several steps power iteration before executing Arnoldi procedure just play a role of improving the quality of starting vector v_0.

3 Numerical Results

In this section, we present the numerical experiments to illustrate the efficiency of accelerating the computation of PageRank vector with Arnoldi method.

In the following numerical experiments, we test four methods for computing PageRank problem, Power method (shorten as 'Power'), Power method with quadratic extrapolation acceleration (shorten as 'QE-Power'), Arnoldi-type method (shorten as 'Arnoldi-type') and Power method with Arnoldi acceleration (shorten as 'Ar-Power'), for dampling factor α varies from 0.85 to 0.99.

In Table 1, we list the characteristics of test matrices including matrix size (n), number of nonzeros (nnz) and density (den), which is defined by

$$\text{den} = \frac{\text{nnz}}{\text{n} \times \text{n}} \times 100.$$

Table 1. The characteristics of test matrices

	matrix	n	nnz	den
1	Globalization	4334	17424	9.28×10^{-2}
2	Recipes	5243	18152	6.60×10^{-2}
3	Search engines	11659	292236	2.15×10^{-1}

The initial vector is $q_0 = (1, 1, \ldots, 1)^T$ and we take $p = 15$ and $m = 6$ in every loop. Since the 1-norm is a natural choice for the Pagerank computation, we choose the 1-norm of the residual as stoping criterion for a fair comparison, i.e.,

$$\|Aq - q\|_1 \leq 1.0 \times 10^{-6}.$$

Numerical experiments were done on a computer with 2.10 GHz CPU and 3G memory.

In Table 2 we list the number of iterative steps and CPU time of the above four methods for different test matrices with the damping factors $\alpha = 0.85, 0.90, 0.95,$ and 0.99 respectively.

From Table 2, it is obvious that the Power method with Arnoldi acceleration is the best approach in terms of both computational time and iteration steps among all the four methods.

It is observed that with the increasing of the damping factor α, the convergence speed of the new method is much faster than the other three methods, especially when the damping factor is close to 1. For test matrix 'Search Engines', when $\alpha = 0.99$, Power method with quadratic extrapolation needs about $4.41s$ to satisfy the convergence rule, but the method we proposed requires only $1.67s$ to reach the same accuracy.

In Fig. 1, we plot the curves of the norm of the residual versus the number of iteration, for the Power method, Power method with quadratic extrapola-

Table 2. The number of iteration steps and CPU time of four methods for test matrices with different α

Matrix	α	Power		QE-Power		Arnoldi		Ar-Power	
		IT	Time	IT	Time	IT	Time	IT	Time
1	0.85	83	0.20	78	0.14	39	0.39	28	0.09
	0.90	129	0.23	110	0.19	45	0.47	34	0.11
	0.95	259	0.39	176	0.28	56	0.56	42	0.17
	0.99	1163	1.64	411	0.66	100	1.03	47	0.19
2	0.85	86	0.19	81	0.16	39	0.50	32	0.13
	0.90	126	0.27	117	0.22	50	0.64	38	0.14
	0.95	256	0.41	201	0.39	68	0.83	42	0.23
	0.99	1191	1.77	469	0.94	126	1.66	60	0.27
3	0.85	81	1.63	40	0.78	32	1.31	28	0.69
	0.90	125	2.38	60	1.19	40	1.63	35	0.83
	0.95	251	4.75	104	2.09	55	2.28	48	1.22
	0.99	1101	20.83	217	4.41	124	5.06	65	1.67

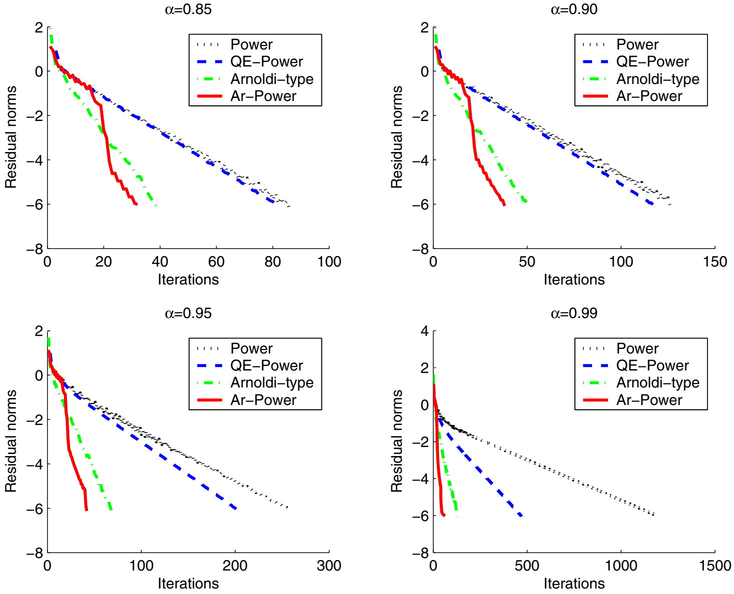


Fig. 1. The norm of residual versus iteration number for test matrix ’Recipes’

tion, Arnoldi-type method and Power method with Arnoldi acceleration for test matrix ’Recipes’ when $\alpha = 0.85, 0.90, 0.95$ and 0.99 respectively.

It is clear from Fig. 1 that our new approach is efficient and can speed up the convergence performance of Power method dramatically.

4 Conclusion

In this paper, we present Power method with Arnoldi acceleration for the computation of Pagerank vector. Our new method is easy to be implemented and can omit complex arithmetic. Numerical results illustrated that our new method is efficient and works better than its counterparts.

In the future, it is of interest to further study the convergence and theoretical property of our new method. In addition, it is also a great challenge to consider the parallel version of this kind of method for large scales Pagerank computation.

References

1. Arnoldi, W.E.: The principle of minimized iteration in the solution of the matrix eigenvalue problem. Quart. Appl. Math. 9, 17–29 (1951)
2. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web, Standford Digital Libraries Working Paper (1998)
3. Berkhin, P.: A survey on PageRank computing. Internet Mathematics 2, 73–120 (2005)

4. Golub, G.H., Greif, C.: An Arnoldi-type algorithm for computing PageRank. BIT Numerical Mathematics 46, 759–771 (2006)
5. Langville, A.N., Meyer, C.D.: Deeper inside PageRank. Internet Mathematics 1, 335–380 (2005)
6. Kamvar, S.D., Haveliwala, T.H., Golub, G.H.: Adaptive methods for the computation of PageRank. Linear Algebra Appl. 386, 51–65 (2004)
7. Kamvar, S.D., Haveliwala, T.H., Manning, C.D., Golub, G.H.: Extrapolation methods for accelerating PageRank computations. In: Proceedings of the Twelfth International World Wide Web Conference (2003)
8. Kamvar, S.D., Haveliwala, T.H., Manning, C.D., Golub, G.H.: Exploiting the block structure of the web for computing PageRank. Stanford University Technical Report, SCCM-03-02 (2003)
9. Haveliwala, T.H., Kamvar, S.D.: The second eigenvalue of the Google matrix. In: Proceedings of the Twelfth International World Wide Web Conference (2003)
10. Wu, G., Wei, Y.: A Power–Arnoldi algorithm for computing PageRank. Numer. Linear Algebra Appl. 14, 521–546 (2007)
11. Sleijpen, G.L.G., Van der Vorst, H.A.: A Jacobi-Davidson iteration method for linear eigenvalue problems. SIAM J. Mat. Anal. Appl. 17, 401–425 (1996)
12. Bellalij, M., Saad, Y., Sadok, H.: On the convergence of the Arnoldi process for eigenvalue problems. Report umsi-2007-12, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN (2007)

A Framework for Automatic Query Expansion

Hazra Imran[1] and Aditi Sharan[2]

[1] Department of Computer Science, Jamia Hamdard, New Delhi, India
[2] School of Computers and System Sciences, Jawaharlal Nehru University, New Delhi, India
himran@jamiahamdard.ac.in, aditisharan@mail.jnu.ac.in

Abstract. The objective of this paper is to provide a framework and computational model for automatic query expansion using psuedo relevance feedback. We expect that our model can be helpful in dealing with many important aspects in automatic query expansion in an efficient way. We have performed experiments based on our model using TREC data set. Results are encouraging as they indicate improvement in retrieval efficiency after applying query expansion.

Keywords: Automatic Query Expansion (AQE), Pseudo Relevance Feedback (PRF), Information Retrieval (IR).

1 Introduction

In an information retrieval system, the query expansion is defined as an elaboration process of user's information need. Reformulation of the user queries is a common technique in information retrieval to cover the gap between the original user query and his need of information. Query expansion is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance. Efthimiadis [5] has done a complete review on the classical techniques of query expansion. Some of the important questions regarding query expansion s are: **What is the source of selecting expansion terms? Given the source, which terms should be selected for expansion? How should weights of terms be calculated?** Source of selecting the terms can be external or internal (from corpus itself). Considering corpus as source of selection, terms can be selected either globally (form entire corpus) or locally (from a subset of documents deemed to be relevant to query). Global analysis methods are computationally very expensive and their effectiveness is generally not better (sometimes worse) than local analysis. In local analysis user may be asked to select relevant documents from set of documents retrieved by information retrieval system. The problem with local analysis is that user's involvement makes it difficult to develop automatic methods for query expansion. To avoid this problem **pseudo relevance feedback (PRF)** approach is preferred in local analysis, where documents are retrieved using an efficient matching function and top n retrieved documents are assumed to be relevant. Automatic query expansion refers to techniques that modify a query without user control. One argument in favor of automatic query expansion is that the system has access to more statistical information on the relative utility of expansion terms and can make a better selection of which terms to add to the user's query. We have worked on local method for automatic query expansion using pseudo relevance feedback. In our previous work [7] we have

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 386–393, 2010.

focused on how thesaurus can be used for query expansion in selecting the terms externally.

In this paper we have proposed a framework for corpus based automatic query expansion. The paper is divided in V sections. In section II, we present a review of related work. Section III describes our proposed framework. Experimental results are presented in Section IV. Section V summarizes the main conclusions of this work.

2 Related Work

Early work of Maron [10] demonstrated the potential of term co-occurrence data for the identification of query term variants. Lesk[8] noted that query expansion led to the greatest improvement in performance, when the original query gave reasonable retrieval results, whereas, expansion was less effective when the original query had performed badly. Sparck Jones [14] has conducted the extended series of experiments on the ZOO-document subset of the Cranfield test collection. Sparck Jones results suggested that the expansion could improve the effectiveness of a best match searching.This improvement in performance was challenged by Minker et al. [11]. More recent work on query expansion has been based on probabilistic models [2]. Voorhees [4] expanded queries using a combination of synonyms, hypernyms and hyponyms manually selected from WordNet, and achieved limited improvement on short queries. Stairmand[9] used WordNet for query expansion, but they concluded that the improvement was restricted by the coverage of the WordNet. More recent studies focused on combining the information from both co-occurrence-based and handcrafted thesauri [13]. Carmel [3] measures the overlap of retrieved documents between using the individual term and the full query. Ruch et al.[12] studied the problem in the domain of biology literature and proposed an argumentative feedback approach, where expanded terms are selected from only sentences classified into one of four disjunct argumentative categories. Cao [6] uses a supervised learning method for selecting good expansion terms from a number of candidate terms.

3 Proposed Framework

Improving robustness of query expansion has been goal of researchers in last few years. We propose a framework for automatic query expansion using Pseudo Relevance feedback. Our framework allows to experiment on various parameters for automatic query expansion. Figure 1 depicts the components in our proposed framework. For a given query, *Information Retrieval system* will fetch top N documents from the corpus similar to the query, based on some similarity measures such as jaccard and okapi similarity measure. *Summary Generation System* takes the ranked Top N documents as input and generates the summary corresponding to each document. Either top N documents or their corresponding summaries will act as *Source for Selecting Expansion Terms*.

Once the source is selected, our next module is for extracting expansion terms. We have used two *Methods for extracting terms*: - one is based on Term co-occurrences and other is based on Lexical Links. Based on term co-occurrence method, we derive expansion terms that are statistically co-occurring with the given query. For our

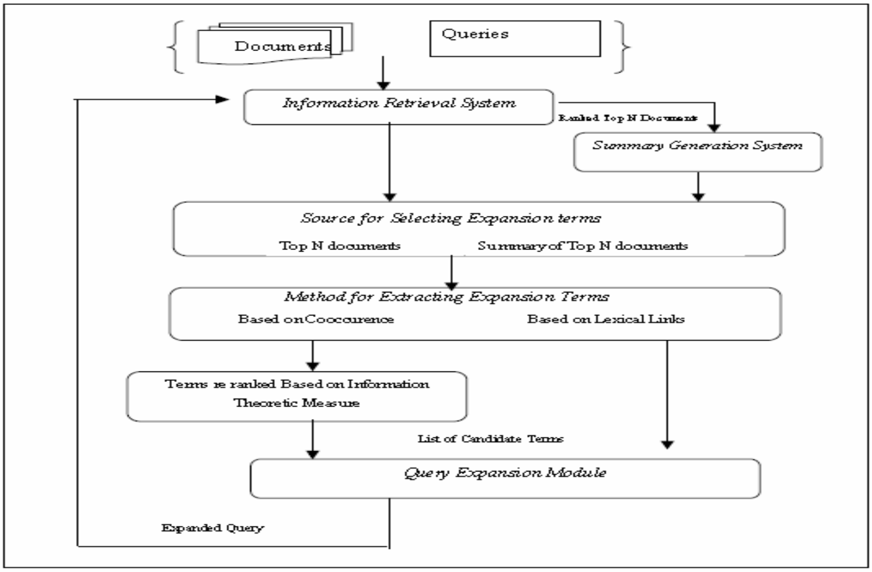

Documents
Queries
Information Retrieval System
Ranked Top N Documents
Summary Generation System
Source for Selecting Expansion terms
Top N documents
Summary of Top N documents
Method for Extracting Expansion Terms
Based on Coocurence
Based on Lexical Links
Terms re ranked Based on Information Theoretic Measure
List of Candidate Terms
Query Expansion Module
Expanded Query

Fig. 1. Proposed automatic query expansion and retrieval framework

experiments, we have used jaccard coefficient for measure the similarity between query terms and all other terms present in relevant documents.

$$jaccard\_co(t_i, t_j) = \frac{d_{ij}}{d_i + d_j - d_{ij}} \tag{1}$$

Where

d_i and d_j are the number of documents in which terms t_i and t_j occur, respectively, and d_{ij} is the number of documents in which t_i and t_j co-occur.

However there is a danger in adding these terms directly the query. The candidate terms selected for expansion could co-occur with the original query terms in the documents (top n relevant) by chance. The higher its degree is in whole corpus, the more likely it is that candidate term co-occurs with query terms by chance. Keeping this factor in mind inverse document frequency of a term can be used along with above discussed similarity measures to scale down the effect of chance factor. Incorporating inverse document frequency and applying normalization define degree of co-occurrence of a candidate term with a query term as follows:

$$co\_\deg ree(c, t_j) = \log_{10}(co(c, t_j) + 1) * (idf(c) / \log_{10}(D)) \tag{2}$$

$$idf(c) = \log_{10}(N / N_c) \tag{3}$$

Where

N = number of documents in the corpus
D= number of top ranked documents used
c =candidate term listed for query expansion

N_c = number of documents in the corpus that contain c
co(c,t_j) = number of co-occurrences between c and t_j in the top ranked documents i.e. jaccard_co(c_i,t_j)

To obtain a value measuring how good c is for whole query Q, we need to combine its degrees of co-occurrence with all individual original query terms. So we use suitability for Q to compute $t_1, t_2, \ldots t_n$

$$SuitabiltyforQ = f(c,Q) = \prod_{t_i in Q} (\delta + co\_\deg ree(c,t_i))^{idf(t_i)} \tag{4}$$

Above equation provides a suitability score for ranking the terms co-occurring with entire query. Still there are chances that a term that is frequent in top n relevant documents is also frequent in entire collection. In fact this term is not a good for expansion, as it will not allow discriminating between relevant and non-relevant document. Keeping this as motivation we suggest the use of information theoretic measures for selecting good expansion terms. We then rank the expansion terms based on the KLD. This approach is based on studying the difference between the term distribution in the whole collection and in the subsets of documents that are relevant to the query, in order to, discriminate between good expansion terms and poor expansion term. Terms closely related to those of the original query are expected to be more frequent in the top ranked set of documents retrieved with the original query than in other subsets of the collection or entire collection. We used the concept of Kullback-Liebler Divergence to compute the divergence between the probability distributions of terms in the whole collection and in the top ranked documents obtained using the original user query. The most likely terms to expand the query are those with a high probability in the top ranked set and low probability in the whole collection. For the term t this divergence is:

$$KLD(t) = [p_R(t) - p_C(t)]\log\frac{p_R(t)}{p_C(t)} \tag{5}$$

where $P_R(t)$ be the probability of t estimated from the corpus R, $P_C(t)$ is the probability of t $\in$V estimated using the whole collection. To estimate $P_C(t)$, we used the ratio between the frequency of t in C and the number of terms in C.

$$P_C(t) = \frac{f(t)}{N} \tag{6}$$

$$P_R(t) = \begin{cases} \gamma\ \frac{f_R(t)}{NR} ift \in v(R) \\ \delta\ p_c(t) otherwise \end{cases} \tag{7}$$

Where C is the set of all documents in the collection, R is the set of top retrieved documents relative to a query, *NR* is the number of terms in R ,v(R) be the vocabulary of all the terms in R, $f_R(t)$ is the frequency of t in R ,f(t)=frequency of term in C,N is the number of terms in C.

The candidate terms were ranked by using equation (7) with γ=1, which amounts to restricting the candidate set to the terms contained in R. Terms not present in relevant sent are given a default probability. The other method is based on Lexical links. The method relies on calculating lexical cohesion between query terms' contexts in a document [15]. The main goal of this method is to rank documents by taking into consideration how cohesive the contextual environments of distinct query terms are in each document. The assumption is that if there is a high degree of lexical cohesion between the contexts of distinct query terms in a document, they are likely to be topically related, and there is a greater chance that this document is relevant to the user's query. Finally *query expansion module* reformulates the query by adding potential candidate terms with the initial query. We have given the weights to expanded query terms using their tf X idf values. The document collection is then ranked against the reformulated query.

4 Experiments

We have used the Vector Space Model implementation to build our information retrieval system. Stemming and stop word removing has been applied in indexing and expansion process. For our experiments, we used volume 1 of the *TIPSTER* document collection, a standard test collection in the IR community. We have used WSJ corpus, and TREC topic set, with 50 topics, of which we only used the title (of 2.3 average word length) for formulating the query. We have used different measures to evaluate each method. The measures considered have been MAP (Mean Average Precision), Precision@5, and Precision @10.

Parameters for performing Automatic Query Expansion using Pseudo Relevance Feedback

Firstly, we investigate the parameters for performing AQE having effect on retrieval performance. The parameters of query expansion are: *Top N doc*↼number of top-ranked documents to be considered as the pseudo-relevance set), *Number of expansion terms*↼the number of informative terms to be added to the query).

1. Number of top ranked documents

Based on the fact that the density of relevant documents is higher for the top-ranked documents, one might think that the fewer the number of documents considered for expansion, the better the retrieval performance. However, this was not the case. As shown in Table1, the retrieval performance was found to increase as the number of documents increased, at least for a small number of documents, and then it gradually dropped as more documents were selected.

This behavior can be explained considering that the percentage of truly relevant documents in the pseudo-relevant documents is not the only factor affecting

Table 1. Performance versus number of pseudo-relevant documents for TREC-1

	5	10	15	20	25	30
Mean Average Precision	0.129902	0.129906	0.129910	0.129696	0.1295	0.12924
PREC-AT-5	0.1344333	0.1342334	0.1344334	0.1344334	0.13243	0.13243
PREC-AT-10	0.1201	0.1202	0.1211698	0.12018	0.12019	0.1191698

performance here. The optimal choice should represent a compromise between the maximization of the percentage of relevant documents and the presence of at least some relevant document. Consistently with the results reported in Table 1, we found that these two parameters were best balanced when the size of the training set ranged from 6 to 15. Further experiments revealed that the system performance decreased nearly monotonically as the number of documents was increased beyond those shown in Table 1. The decline in performance was however slow, because the number of relevant documents remained substantially high even after a large number of retrieved documents. For instance, for TREC-1, the average precision at 20 documents was 0.129696, at 30 documents was 0.12924.

2. Number of terms selected for expansion

This section seeks to answer questions regarding the best number of query terms to use for expansion. We let the number of expansion terms vary from 5 to 30 (step = 5), computing for each value the retrieval performance of the system. Table 2 shows that the maximum values of the different performance measure were reached for different choices of the number of selected terms. Most important, the results show that the variations in performance were negligible for all measures and for all selected sets of expansion terms. With respect to *number of expansion terms* considered in the QE, using less than 10 terms means a drop-off in MAP, while for *number of expansion terms* ≥ 10, the retrieval performance is stable. To assess the stability of the approaches with respect to *number of* expansion term and top N doc, we vary them and record the MAP. In particular, we vary 2 ≤ top N doc ≤ 30 and 1 ≤ number of expansion term ≤ 20.

Table 2. Performance versus number expansion terms for TREC-1

	5	10	15	20
Mean Average Precision	0.129906	0.132393	0.139817	0.139802
PREC-AT-5	.1330334	0.1342234	0.1344334	0.1341334
PREC-AT-10	0.1172	0.1200169	0.1211698	0.1211698

Figure 2 present surface plots of Query Expansion Settings for KLD.

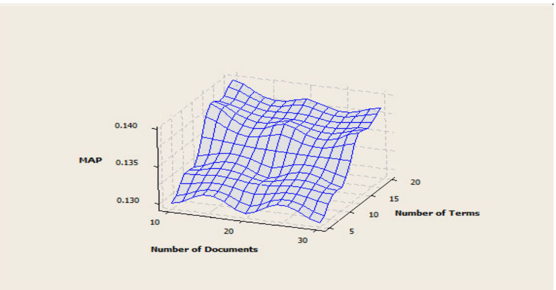


Fig. 2. Surface plots of MAP for Query Expansion when the number of Top N documents and number of Expansion Terms parameters are varied

In our framework, we have used two sources for selecting expansion terms. One is a top N document and other is summary of top N documents. Again, we suggest two methods for extracting terms .One is based on co-occurrence and other is on lexical links. Figure 3 shows the graph of an example of a query where terms are extracted based on term co-occurrence , KLD and lexical information respectively on different queries.

Analysis of result shows that after intensive computation we may select appropriate parameters for AQE . Further, we observe that re-ranking co-occurring terms on KLD, improve result significantly. For method based on lexical link we observe that sufficient links are available only for few queries. However, for the queries where sufficient links are found, retrieval performances improved in most of the cases.

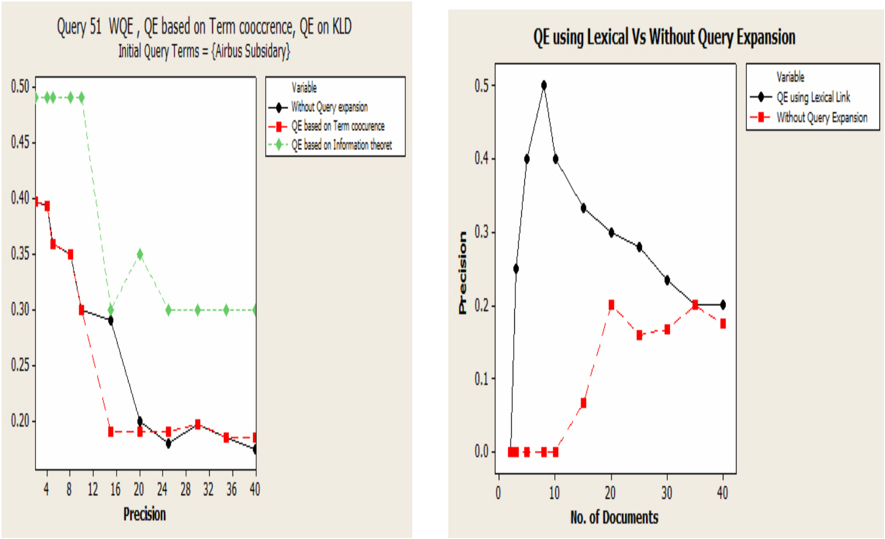


Fig. 3. Graph of a particular query for without query expansion, QE based on Term co-occurrence and information theoretic measure and Lexical Information

5 Conclusion

In this paper we have proposed a framework along with a computational model for automatic query expansion using PRF. Our framework is flexible and feasible. Regarding flexibility, it allows you to experiment with different methods for selecting top n relevant, selecting query expansion terms, and selecting parameters for query expansion. Regarding feasibility it provides step-by-step procedure for implementing query expansion. Analysis of our results shows that query terms selected on the basis of co-occurrence are related to original query, but may not be good discriminator to discriminate between relevant and non-relevant document. KLD measure allows this discrimination to certain extent; hence it improves retrieval performance over co-occurrence based measure. Lexical links allows us to deal with the context of query in addition to co-occurrence measure. We are exploring use of semantic links for improving lexical based query expansion.

References

1. Lee, C.J., Lin, Y.C., Chen, R.C., Cheng, P.J.: Selecting effective terms for query formulation. In: Proc. of the Fifth Asia Information Retrieval Symposium (2009)
2. Croft, W.B., Harper, D.J.: Using probabilistic models of document retrieval without relevance information. Journal of Documentation 35, 285–295 (1979)
3. Carmel, D., Yom-Tov, E., Soboroff, I.: SIGIR Workshop Report: Predicting query difficulty – methods and applications. In: Proc. of the ACM SIGIR 2005 Workshop on Predicting Query Difficulty – Methods and Applications, pp. 25–28 (2005)
4. Voorhees, E.M.: Query expansion using lexical semantic relations. In: Proceedings of the 1994 ACM SIGIR Conference on Research and Development in Information Retrieval (1994)
5. Efthimiadis, E.N.: Query expansion. Annual Review of Information Systems and Technology 31, 121–187 (1996)
6. Cao, G., Nie, J.Y., Gao, J.F., Robertson, S.: Selecting good expansion terms for pseudorelevance feedback. In: Proc. of 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243–250 (2008)
7. Imran, H., Sharan, A.: Thesaurus and Query Expansion. International journal of computer science & information Technology (IJCSIT) 1(2), 89–97 (2009)
8. Lesk, M.E.: Word-word associations in document retrieval systems. American Documentation 20, 27–38 (1969)
9. Stairmand, M.A.: Textual context analysis for information retrieval. In: Proceedings of the 1997 ACM SIGIR Conference on Research and Development in Information Retrieval (1997)
10. Maron, M.E., Kuhns, J.K.: On relevance, probabilistic indexing and information retrieval. Journal of rhe ACM 7, 216–244 (1960)
11. Minker, J., Wilson, G.A., Zimmerman, B.H.: Query expansion by the addition of clustered terms for a document retrieval system. Information Storage and Retrieval 8, 329–348 (1972)
12. Ruch, P., Tbahriti, I., Gobeill, J., Aronson, A.R.: Argumentative feedback: A linguistically-motivated term expansion for information retrieval. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pp. 675–682 (2006)
13. Mandala, R., Tokunaga, T., Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion. In: Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval (1999)
14. Sparck Jones, K.: Automatic keyword classification for information retrieval. Butterworth, London (1971)
15. Vechtomova, O., Wang, Y.: A study of the effect of term proximity on query expansion. Journal of Information Science 32(4), 324–333 (2006)

WSCache: A Cache Based Content-Aware Approach of Web Service Discovery*

Dunlu Peng, Xinglong Jiang, Wenjie Xu, and Kai Duan

School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai 200093, China
dunlu.peng@gmail, j2910@163.com, xu071602@126.com,
duankai@yeah.net

Abstract. Web service discovery is an important issue for the construction of service based architectures and is also a prerequisite for service oriented applications. Nowadays, Data-intensive Web Service (DWS) is widely used by enterprises to provide worldwide data query services. Different from the existing service discovery approaches, in this work, we propose a service discovery mode for DWS which combines the characteristics of DWS with semantic similarity of services. The model, named as RVBDM, discovers services based on the return values of DWS. Some algorithms, such as Web service cache Update (WSCU) algorithm and Web Service Match (WSM) algorithm, are presented to implement the model. Experiments were conducted to verify the performance of our proposed approach.

1 Introduction

Nowadays, Web service becomes an important Internet technology. Web service ensures the interoperation of applications on different platforms. Technologies used in Web service are based on open standards protocol specification, such as XML, DWSL and SOAP protocol. With the development of Web services ,how to find the proper Web services for requestors has become a highlight in research area. This issue, which is specified as Web service discovery, is very important for the implementation of service based systems [1].

Data-intensive Web Service (DWS) [3], which is supported by hidden database, is widely applied by enterprises in recent years. Its main goal is to provide data query for its requestors. In existing approaches of Web service discovery, the requestor searches WSDL in UDDI to check whether there are proper services. This kind of approaches is based on keyword matching and easy to use. However, its performance can’t be guaranteed for the lack of taking the content of service into account during the discovery. In this work, we combine the DWS characteristics with the semantic similarity of DWS and propose a cache based content-aware approach for DWS discovery.

* This work is support by Municipal Nature Science Foundation of Shanghai under Grant No. 10ZR1421100, Innovation Program of Shanghai Education Commission under Grant No. 08YZ98, Postgraduate Construction of the Core Curriculum of University of Shanghai for Science and Technology under Grant No. 16000206 and Innovation Program of School of Optical-Electrical and Computer Engineering, USST, under Grant No. GDCX-T-102.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 394–401, 2010.

Our work is organized as follows: In Section 2, we discus the model of DWS discovery; the framework of service cache is described in Section 3; The algorithms for updating service cache and searching of service are presented in Section 4; In Section 5, we investigate the experiments which are conducted to verify the performance of our approaches and finally we draw the conclusion in Section 6.

2 Model of DWS Discovery

In this section, we introduce the definition of semantic similarity and specify the principle and structure of RVBDM model.

2.1 Semantic Similarity

Web service discovery is the process of finding the best match demand services in accordance with the functional requirements provided by the service requester. At present, the main technologies of Web service discovery are classified into three categories [1]. The first one is Keywords-based approaches [4]. Another one is Frame-based approaches, are lack of semantic support of UDDI and have low precision and recall [7,8]. The last one is Semantic-based approaches. They use semantic information and ontology effectively and match services at semantic level [2].

The similarity of two data-intensive Web services can be determined with semantic similarity of return values. According to these return values, we can judge whether a Web service meets requestor's request. The definition of semantic similarity of two data-intensive Web services is given as follows:

***Definition 1 (Pairwised* Semantic Similarity).** *Given two data-intensive Web services, whose return values are grouped into two concepts* $C_1=\{c_{11},c_{12},\cdots,c_{1n}\}$ *and* $C_2=\{c_{21},c_{22},\cdots,c_{2m}\}$*, the* **pairwised semantic similarity** *of these two services, denoted as* $SemSim(C_1,C_2)$*, is computed using following formula:*

$$SemSim\ (C_1,C_2) = Max\left(\sum_{i=1}^{n}\sum_{j=1}^{m} sim\ (c_{1i},c_{2j})\right) / Min\ (n,m) \tag{1}$$

where $Sim(c_{1i},c_{2j})$ is the semantic similarity between two categories of concepts, which includes the similarity of syntax level *SimT*, the similarity of internal structure *SimIS*, the similarity of external structure *SimES*, as well as the similarity of extension *SimEX*. $Sim(c_{1i},c_{2j})$ is the sum of these four weighted value: $Sim(c_{1i}, c_{2j})= \omega_1 SimT(c_{1i}, c_{2j})+ \omega_2 SimIS(c_{1i}, c_{2j})+ \omega_3 SimES(c_{1i},c_{2j})+ \omega_4 SimEX(c_{1i},c_{2j})$, in which $\omega_1+\omega_2+\omega_3+\omega_4=1.0$. Where *SimT* is the similarity based on Longest Common Substring; *SimIS* is to compute similarity by comparing property of concepts; *SimES* is to compute similarity by comparing concepts' super-property and sub-property; *SimEX* is to get similarity by comparing similarity of concept's instance.

2.2 Services Discovery Model

Based on the theory of semantic similarity, we propose the services discovery model RVBDM (Return Value-Based Discovery Model, RVBDM) which is shown in Fig. 1.

We describe the process of RVBDM to discover Web Services as follows:

Step 1: RVBDM accepts requestor's demand.

Step 2: Matching calculator determines whether there are some services fulfilling requestor's demand, in service cache. If existing, forward to 3), else forward to 4).

Step 3: Return the services to requestor. If the requestor is not satisfied with current result, he/she can submit another request and forward to 4).

Step 4: Go to UDDI directly to search services. If RVBDM finds proper services, it returns them to the requestor. Otherwise, it reports matching failure.

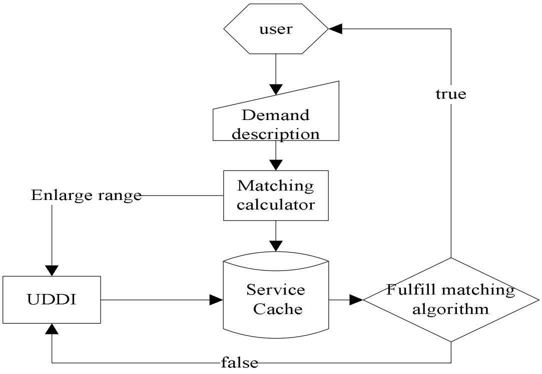


Fig. 1. Sketch of RVBDM

3 Service Cache

We now present the structure of service cache, as well as the method for initializing and updating the cache.

3.1 Architecture of Service Cache

Service cache is an important component of RVBDM.A list of available Web services are saved in it. Service cache assigns each of these Web services an identifier WSID, and stores information for each Web service, such as its URI, return value, functional description, and clicks etc. Service cache consists of following three components (see Fig. 2):

1. Return Value Clusters (RVC). A RVC is a collection of return values. These return values meet the same requirements on semantic similarity. RVBDM extracts the Eigen value for each RVC $\lambda_1 ,\lambda_2 ,\lambda_n$. Elements in each RVC have the same semantic similarity on λ_n or located in a certain range.
2. Service Description Clusters (SDC). A SDC is a collection of service description. These descriptions meet the same requirements on semantic similarity. SDCs are classified in the same way as RVCs.
3. Mapping Service description with corresponding service. With different classification criterion, each Web service in the service list may be mapped into more than one SDC.

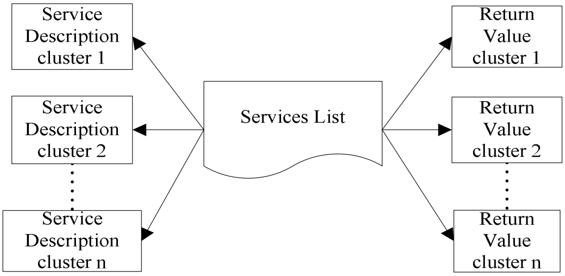

Service Description cluster 1
Service Description cluster 2
Service Description cluster n
Services List
Return Value cluster 1
Return Value cluster 2
Return Value cluster n

Fig. 2. Architecture of service cache

All of the above components are associated with each other by WSID. Their relationship is employed in our Web service matching algorithm.

3.2 Initializing and Updating Service Cache

Initially, in the RVBDM, service list is set as an empty table. Two events trigger the update of service cache:

1. RVBDM periodically searches the available services in UDDI which do not exist in the service cache. Then, RVBDM sends a tentative request to the service and saves the return value into the service list.
2. The other event is requestor's request. When requestor is not satisfied with the returned services or there is no service satisfying the request in service cache.

When update is performed, RVBDM creates RVC and SDC in service cache according to the return values and description of the service.

3.3 Matching Calculator

Matching calculator is another important component of RVBDM. Its functionality is to filter Web services that fulfill requestor's demand from service cache or UDDI, and rank the result simply. The algorithm used by matching calculator is introduced in the next section.

4 Algorithms

We present the algorithms and theory of RVBDM for searching Web services in this section.

4.1 Service Cache Update Algorithm-WSCU

There are two main operations for updating service cache: one is to put the return value (*RV*) into an appropriate RVC and the other is to extract the key word of service description (*SD*) and put it into an appropriate SDC.

Denote each RVC as a keywords set $\lambda = \{ RV_1, W_1; RV_2, W_{2;} \ldots RV_m ,W_m \}$, RV_m is one return value of the Web service, W_m is the appearing time of return value computed in *WSCU* algorithm. Simply we denote one return values of *S* with a special keyword set $\lambda'=\{RV, 1\}$. This set has only one keyword *RV*, the appearing times of keyword is 1. *SR* is put into RVC λ if similarity between λ and λ' is more than or equal the threshold θ.

We extract the keyword of *SD* using the approach proposed in reference [5] before the operation. Some specific keywords are extracted from *SD* and denote them as $T=\{T_1, T_2,,\ldots T_i\}$. Generally, the coverage and specificity is enough when 10 to 15 keywords are chosen randomly from *SD* [6]. *WSCU* algorithm has little difference with traditional TF/IDF model, see Formula 2:

$$TFIDF\quad(T_i, D_j) = TF\ (T_i, D_j) = \frac{C_{Ti}}{C_{SD}} \tag{2}$$

in which C_{Ti} is the appearing times of keyword T_i in *SD* and C_{SD} is the total number of keywords in *SD*. We choose the first 3-5 words after sorting T_i in descend order according to the value of $TFIDF(T_i, D_j)$ .For simplification, we choose only the first keyword in the description of *S*. After preparation, we classify *SD* in the same way as *RV* classification and put services into appropriate *SDC* according to the keywords of their description. Fig.3 gives the description of the algorithm *WSCU.*

4.2 Algorithm for Web Service Matching (WSM)

The *WSM* algorithm is described in Fig.3. When a search request is submitted, RVBDM matches the return values first (Procedure *firstMatch*), if the return values are similar, match of the description will be performed (Procedure *secondMatch*). If match of return values gives a negative result, RVBDM will continue to search the Web services not cached in service cache (Procedure *findFromOuter*). According to Formula 1, we view the return value of target Web service as concept C_1 and view the return value in requestor's request as concepts C_2. The return values in requestor's request are match with return value of Web service one by one, and there is only one element in C_2. Therefore, we simplify Formula 1 as follows:

$$SemSim\quad(C_1, C_2) = Max\left(\sum_{i=1}^{n} sim\ (c_{1i}, c_2)\right) \tag{3}$$

```
Algorithm. WSM(Web Service Matching)
Input: r-service searching requirements;
         $\lambda$-the vector of r, $\lambda_i$ is the ith component of $\lambda$;
         $\theta$-the threshold of similarity;
Result: RC[]-the appropriate clusters for w
Process:
         For(i=1 to RVC.size){//1 first matching for RVC
            I f(RVC[i] contains the unit of measurement)
            Ui ←RVC[i]; //extract the unit of measurement
            R←firstMatch (Ui, $\lambda\mu$[],$\theta$);
          Else
                  R←firstMatch(RVC [i], $\lambda\mu$[],$\theta$);
          End IF
           End For
```

Fig. 3. Algorithm of Web Service Matching (WSM)

```
  R' ←secondMatch();//2 second matching for SDC
    FR←R∪R';//merge R and R'
    FR←rank(FR);//sort the final result FR
    Return FR;
Procedure: firstMatch (r, $\lambda$, $\theta$)
Input: r, $\lambda$, $\theta$-same meaning as Algorithm WSCU;
Process:
          r← $\lambda_i'$ ;//wrap a into vector space
                For(j=1 to RVC.size)
            If (Sim( $\lambda_i'$ ,RVC[j])≥ $\theta$)
         RC← RVC[j]; //put $\lambda_j$ into the result set
         flag ← true;
                 break;
              End If
          End For
         If (!flag)// cannot find appropriate service for r
       RC←findFromOuter(r, $\theta$);
         End if
        Retrun RC;
Procedure: findFromOuter(r, θ)
Input: r, $\theta$-same meaning as Algorithm WSCU;
Process:
          counter ←1;
    While (true)
       C1 ←RVC;// wrap RVC into concept C1
         If(SemSim(C1,r)≥$\theta$)
         ServiceList←C1;//put S into Service List
                 RC←C1 ;
         End IF
      If(counter≥n)
                break;
   End IF
        End While
Procedure: secondMatch( )
Input: nothing
Process:
          For(i←1≤SDC.size;i++){
              WSIDi ← SDC[i].WSID;//extract every WSID in R
                     For(j←1;j≤SDC[i].size)
      If(WSIDi in SDC[j]){//if SD contains this WSID
            R' ←SDC[j];//put all the Service of SD into R'
              End IF
      End For
        End for
     Return R'
```

Fig. 3. (*continued*)

5 Experiments

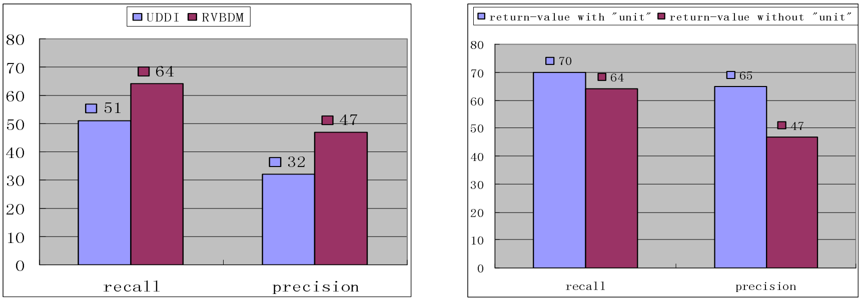
We conducted two sets of experiments to measure RVBDM's precision and recall of service discovery. The first one tests its performance by comparing it with traditional UDDI. The second one evaluates the performance through return values with the unit of measure and without the unit of measure under different similarity threshold.

5.1 Experimental Settings

The experiment simulates the actual running environment of RVBDM on a PC. There are 50 Web services available on it and these Web services are divided into two groups: one is cached in the service cache and the other is not. Both groups contain 10 Web services whose return value has the unit of measurement. Then, using algorithm WSCU, the return value and service description of Web Services are inserted into the corresponding RVC and SDC. In the first matching of WSM algorithm, the similarity threshold θ is set to 0.5, and the maximum number of external Web services found n is set to 2. In addition, the threshold θ1 in WSCU algorithm is set to 0.6.

5.2 Experimental Results

The experimental results of comparison between RVBDM and UDDI are shown in Fig.4. From the figure, we observe that our proposed RVBDM has better performance than traditional UDDI both in precision and recall.

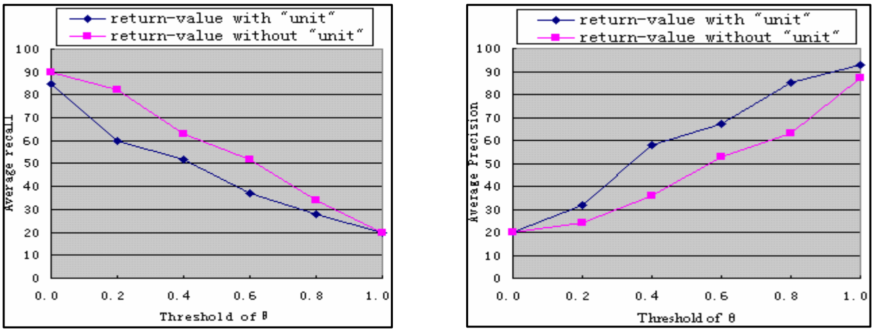


(a) Comparison of RVBDM and UDDI

(b) Comparison of Return Value with/ without the Unit of Measurement

Fig. 4. Comparison Between RVBDM and UDDI

The results of the second group of experiments are shown in Fig.5. The result shows that the precision increases as the increase of similarity threshold, while the recall decreases as the increase of similarity threshold.



(a) Precision (b) Recall

Fig. 5. Effect of Similarity Threshold on Searching Performance

6 Conclusion and Future Work

Service discovery is an important issue for service-oriented application. An effective discovery approach can promote the development of Web applications. In this paper, we propose a Web service discovery model called RVBDM for data-intensive web services which takes into account the Web service's return values when discover services. Some algorithms such as algorithm *WSCU* and *WSM* are designed to implement the model. The experimental results show that our approach is better than the keyword-based discovery. In our future work, we will study how to maintain the service cache more efficiently. And how to choose a proper similarity threshold in *WSCU* algorithm and *WSM* algorithm will be also studied.

References

1. Yue, K., Wang, X., Zhou, A.: Underlying Techniques for Web Service. Journal of Software 15(3), 428–442 (2004) (Chinese)
2. Klein, M., Bernstein, A.: Searching for Services on the Semantic Web Using Process Ontologies. In: Proceedings of the International Semantic Web Working Symposium (SWWS), IOS Press, Amsterdam (2001)
3. Caverlee, J., Liu, L., Rocco, D.: Discovering and Ranking Web Services with BASIL: A Personalized Approach with Biased Focus. In: Proceedings of the 2nd International Conference on Service Oriented Computing (ICSOC 2004), New York, November 15-18 (2004)
4. Dong, X., Halevy, A., Madhavan, J., Nemes, E., Zhang, J.: Similarity Search for Web Services. In: Proceedings of the 30th VLDB Conference, Toronto, Canada (2004)
5. W3C Working Group: Web Services Architecture (February 2004), `http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/`.
6. UDDI org: Universal Description, Discovery, and Integration of Businesses for the Web[EB/OL], `http://www.uddi.org/` (October 1, 2010)
7. Yao, C.J., Miller, S., Qianhui, A., Yang, L.: Service Pattern Discovery of Web Service Mining in Web Service Registry-Repository. In: Proceedings of IEEE International Conference on e-Business Engineering (2006)
8. Paolucci, M., Kawamura, T., Payneand, T.R., Sycara, K.P.: Importing the Semantic Web in UDDI. In: Proceedings of Web Services, E - business and Semantic Web Workshop

A Novel Local Optimization Method for QoS-Aware Web Service Composition

Xiaojie Si[1,2], Xuyun Zhang[1,2], and Wanchun Dou[1,2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] Department of Computer Science and Technology, Nanjing University,
210093, Nanjing, China
mailsxj@126.com, xyzhanggz@gmail.com, douwc@nju.edu.cn

Abstract. QoS-aware web service selection has become a hot-spot research topic in the domain of web service composition. In previous works, the multiple tasks recruited in a composite schema are usually considered of equal importance. However, it is unreasonable for each task to have the absolutely same weight in certain circumstances. Hence, it is a great challenge to mine the weights among different tasks to reflect customers' partial preferences. In view of this challenge, a novel local optimization method is presented in this paper, which is based on a two-hierarchy weight, i.e., weight of task's criteria and weight of tasks. Finally, a case study is demonstrated to validate the feasibility of our proposal.

Keywords: QoS, weight, service composition, local optimization.

1 Introduction

Service-oriented Computing (*SOC*) is a paradigm where the capabilities of network-accessible services can be easily searched and integrated among multiple organizations [1]. Besides, web service technologies are a promising solution to *SOC* [2]. In addition, with the number of web services that share similar functionality increasing, QoS model has been employed to discriminate all these services. Many methods, e.g., global optimization and local optimization have been recruited for the QoS-aware service composition [2]. The local optimization approach, where service selection is done for each task individually, rarely guarantees the global constraints. Our work aims at arguing that the tasks tend to be of different importance in some cases, i.e., weights should be assigned to corresponding tasks according to their importance. Also we propose a novel local optimization method based on weights of two hierarchies, which could guarantee the global constraints and maximize the probability of reaching customers' satisfactory as far as possible.

The remainder of the paper is organized as follows: Section 2 gives the motivation and preliminary knowledge. Section 3 demonstrates our proposed local optimization method. An online shopping case is provided to explain our method more clearly in Section 4. Section 5 discusses the related work. Section 6 concludes the paper.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 402–409, 2010.

2 Motivation and Preliminary Knowledge

Generally, tasks in web services composition process are considered of equal importance, e.g., in [3][4]. However, it is unreasonable for each task to have the absolutely same weights in certain circumstances. Take the online shopping scenario as an instance. Usually, online shopping process consists of three primary steps (see Fig. 1). Assume that there exist two qualified composite solutions, both of which totally cost 40$. The cost of the first consists of 30$ for T-shirt and 10$ for transportation, while that of the second is 35$ and 5$ respectively. To a large extent, the latter solution may be better than the first one, as the customer is prone to spend more money on the T-shirt rather than shipping. Consequently, the task of choosing merchandise is more important than other tasks. In light of this, we argue that different tasks may hold different weights in certain service composition applications. In this situation, it is a challenge that how to mine the task's weights and utilize them for service composition to partially reflect customers' preferences. Here, a framework for QoS-aware web service composition is presented for facilitating our further discussion.

Definition 1. The framework for QoS-aware web service composition is formally represented by a seven-tuple {$\boldsymbol{T}$, $\boldsymbol{SP_i}$, $\boldsymbol{CS}$, $\boldsymbol{C}$, $\boldsymbol{Cons}$, $\boldsymbol{W_i}$, $\boldsymbol{W}$}, whose elements are depicted as below in sequence:

- ***Tasks***, $\boldsymbol{T} = \{t_1, \ldots, t_\mathrm{i}, \ldots, t_l\}$, where t_i $(1 \le i \le l)$ is a task which denotes a abstract service in a composition schema.
- ***Service pool***, $\boldsymbol{SP_i} = \{s_{i1}, \ldots, s_{ij}, \ldots, s_{in}\}$ is a service pool in which s_{ij} $(1 \le j \le n)$ could meet all the functional requirements of t_i.
- ***Composition Schema***, $\boldsymbol{CS} = \{(t_1, \boldsymbol{SP_1}), \ldots, (t_\mathrm{i}, \boldsymbol{SP_\mathrm{i}}), \ldots, (t_\mathrm{l}, \boldsymbol{SP_l})\}$ represents the composition schema that contains l tasks, where (t_i, SP_i) $(1 \le i \le l)$ refers to task t_i as well as its corresponding qualified service pool $\boldsymbol{SP_i}$.
- ***Service Cirteria***, $\boldsymbol{C} = \{c_1, \ldots, c_\mathrm{i}, \ldots, c_m\}$, where c_i $(1 \le i \le m)$ is a QoS criterion of a web service, and $s_{ij}.c_k$ $(1 \le k \le m)$ denotes c_k on the service s_{ij}.
- ***Global Constraints***, ***Cons*** = $\{cons_1, \ldots, cons_i, \ldots, cons_M\}$ refers to a set of global QoS constraints proposed by a customer. Since customers are apt to express their requirements in the form of numeric scopes, so $cons_i$ could be denoted by an interval $[min_i, max_i]$. s_{ij} is qualified for task t_k on $cons_k$, iff the value of $s_{ij}.c_k$ $(1 \le k \le m$) is in the range $[min_k, max_k]$ of the $cons_k$.
- ***Weights of Task Criteria***, $\boldsymbol{W_i} = \{w_{i1}, \ldots, w_{ij}, \ldots, w_{im}\}$, where w_{ij} $(1 \le i \le l$, $1 \le j \le m)$ denotes the weight of constraint c_j $(c_j \in \boldsymbol{C})$ of task t_i.
- ***Weights of Task***, $\boldsymbol{W} = \{w_1, \ldots, w_\mathrm{i}, \ldots, w_\mathrm{l}\}$. w_i $(1 \le i \le l)$ indicates the weight of task t_i and l represents the number of the tasks.

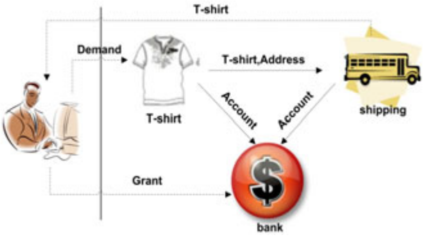


Fig. 1. Online Shopping

Please note that in our paper, we assume the values of m and M are equal. A process-based composite web service is an aggregation of multiple web services, which could interact with each other based on a pre-existed composite solution ***CS*** [4]. The background of our method is the process-based composition. The critical issue of composition is how to choose a suitable service s_{ij} $(1 \le j \le n)$ in $\boldsymbol{SP_i}$ for task t_i $(1 \le i \le l)$, so that the solution not only meets customers' requirements but also is optimized. For simplicity, only the sequential composition model is discussed in this paper. Other models (e.g., parallel, conditional and loops) could be transformed into the sequential one based on [5].

3 A Local Optimization Method Statisfying Global Constrainsts Based on Tasks' Weights

Fig. 2 illustrates the detailed process of our method. The composite schema ***CS*** shown in the left of Fig. 2 consists of three tasks, namely, $\boldsymbol{T} = \{t_1, t_2, t_3\}$. The right section of the figure formally specifies the three steps of our method. Next, we will explicate the specific process of our proposed method.

1) Setp1: Task Weights Calculation

The similarities between the profiles of tasks and user requirements are converted into the weights of tasks. Customers' demand profile is described as an *XML* file (see Fig. 3). According to the context-based matching technique in [6], we employ a string matching approach to calculate the weights based on domain-specific ontology. Context extraction [6] is recruited for the two categories of profiles above. Let $\boldsymbol{U} = \{d_1, \ldots, d_i, \ldots, d_n\}$ denotes a set of extracted descriptors from the customers' demand profile, and $\boldsymbol{T_k} = \{t_{k1}, \ldots, t_{kj}, \ldots, t_{km}\}$ for the extracted descriptors of task t_k. Thus the similarity between T_k and U denoted as:

$$match(T_k,U) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} match_{str}(d_i,t_{kj}) \tag{1}$$

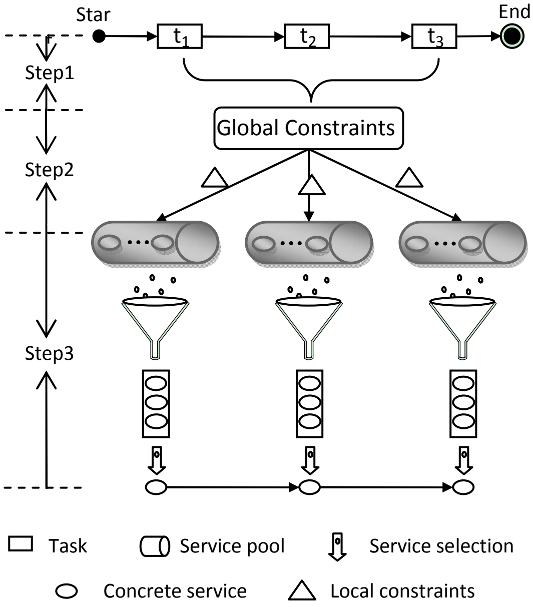
could be calculated by the string-matching function indicated by $match_{str}$ where return 1 iff the u_i match t_i. Finally, according to the following function:

$$weight(t_k) = \frac{match(T_k,U)}{\sum_{i=1}^{l} match(T_i,U)} \tag{2}$$

where l is the number of the tasks, the value of $match(T_k,U)$ is normalized to the weight w_k ($w_k \in W$) for task t_k. Thus equation $\sum_{w_k \in W} w_k = 1$ holds.

2) Setp2: Global Constraints Decomposition

According to the weight set $\boldsymbol{W}$ of tasks, every global constraint $cons_k$ is divided into sub-constraints for each task, so that our local optimization method can meet global constraints. The following formulas are utilized to calculate local constraints of $cons_k$ whose interval value is $[min_k, max_k]$ for the first task t_1, where a set of real variables x_{ik} and y_{ik} are introduced, such that $[x_{ik}, y_{ik}]$ denotes the local constraint of $cons_k$ for task $t_{i.}$



Step1: Task weights calculation. The weights of tasks are calculated through the similarity analysis between the profiles of each task and user requirements.

Step2: Global constraints decomposition. The global constraints proposed by customers are decomposed into local constraints based on the tasks' weights.

Step3: Local optimization selection. According to the local constraints, qualified services are filtered out from the corresponding service pool, and then the service with the high score is selected for composition process.

Fig. 2. The process of our proposed method of local optimization

$$\begin{cases} x_{1k}:\cdots:x_{ik}:\cdots:x_{lk}=w_1:\cdots:w_i:\cdots:w_l \\ F(x_{1k},\cdots,x_{ik},\cdots,x_{lk})=min_k \end{cases} \quad (3) \qquad \begin{cases} y_{1k}:\cdots:y_{ik}:\cdots:y_{lk}=w_1:\cdots:w_i:\cdots:w_l \\ F(y_{1k},\cdots,y_{ik},\cdots,y_{lk})=max_k \end{cases} \quad (4)$$

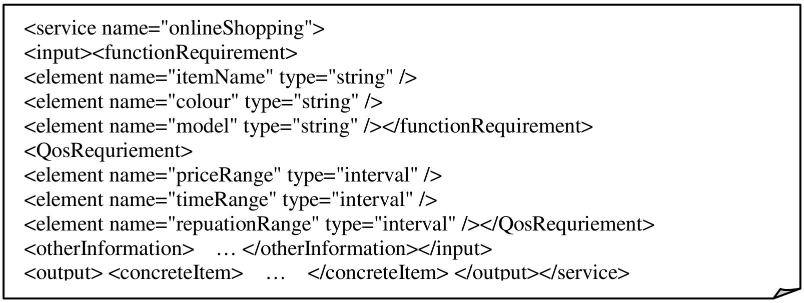
The proportion of a set of variable x_{ik} (y_{ik}) is equal to that of a group of w_i. Function F denotes the aggregation function for $cons_k$. Here, we present the basic function $\boldsymbol{F}$ that could be given by:

$$F(x_{1k},\cdots,x_{ik},\cdots,x_{lk})=g\sum_{i=1}^{l}x_{ik}(g\in R) \quad (5) \qquad F(x_{1k},\cdots,x_{ik},\cdots,x_{lk})=g\prod_{i=1}^{n}x_{ik}(g\in R) \quad (6)$$

where the formula (5) denotes the values ($s_{ij}.c_k$) (e.g., price, duration, reputation) needed to be summed up, and formula (6) represents the values needed to be multiplied (e.g., availability, success rate), which are proposed in [3]. The local constraint of $cons_k$ for task t_{i+1} ($i \geq 1$) is computed as:

$$\begin{cases} x_{(i+1)k}:\cdots:x_{lk}=w_{i+1}:\cdots:w_l(i\geq 1) \\ F(x_{(i+1)k},\cdots,x_{lk})=min_k \\ x_{1k}=s_1.c_k \\ \quad\cdots \\ x_{ik}=s_i.c_k \end{cases} \quad (7) \qquad \begin{cases} y_{(i+1)k}:\cdots:y_{lk}=w_{i+1}:\cdots:w_l(i\geq 1) \\ F(y_{(i+1)k},\cdots,y_{lk})=min_k \\ y_{1k}=s_1.c_k \\ \quad\cdots \\ y_{ik}=s_i.c_k \end{cases} \quad (8)$$

After the service s_i has been selected for task t_i, perhaps the remaining resources that are available could be appropriated for later service selection. As a result, the local constraints for t_{i+1} could possibly be modified. For example, the local constraint price



```
<service name="onlineShopping">
<input><functionRequirement>
<element name="itemName" type="string" />
<element name="colour" type="string" />
<element name="model" type="string" /></functionRequirement>
<QosRequriement>
<element name="priceRange" type="interval" />
<element name="timeRange" type="interval" />
<element name="repuationRange" type="interval" /></QosRequriement>
<otherInformation>   … </otherInformation></input>
<output> <concreteItem>   …   </concreteItem> </output></service>
```

Fig. 3. Customers' demand profile

$cons_{price}$ for t_i is [4\$, 10\$] and the selected service $s_{ij}.c_{price}$ is 6\$, then 4\$ can be utilized for the task t_{i+1}. Therefore, when the local constraint of $cons_k$ for task t_{i+1} is calculated, the values of QoS criteria of these selected services from t_1 to t_i should be considered. Finally, the interval $[x_{(i+1)k}, y_{(i+1)k}]$ denotes the local constraint of $cons_k$ for task t_{i+1}.

3) Step3: Local Optimization Selection

The service selection is the core of our method. Briefly speaking, it is to select a service s_i that meets the local constraints and owns the highest score in $\boldsymbol{SP}_i$ for task t_i. Because different criteria own different units and consist of both positive and negative criteria, all the QoS criteria are scaled and integrated into a unified value through the *SAW* technique [7]. For negative criteria, when the value gets higher, the quality of service tends to be lower. However, the case for positive criteria is completely contrary. For negative criteria, formula (9) is used to scale the value, while formula (10) is employed to scale the positive criteria.

$$value' = \left\{ \frac{Qmax(s_i.c_k, y_{ik}) - value}{Q\max(s_i.c_k, y_{ik}) - Q\min(s_i.c_k, x_{ik})} \right. \tag{9}$$

$$value' = \left\{ \frac{value - Qmin(s_i.c_k, x_{ik})}{Qmax(s_i.c_k, y_{ik}) - Qmin(s_i.c_k, x_{ik})} \right. \tag{10}$$

In the equations above, *value* represents $s_{ij}.c_k$, x_{ik} or y_{ik}. *Value*' denotes the scaled value($s_{ij}.c_k'$, x_{ik}' or y_{ik}'). For negative criteria, x_{ik}' is greater than y_{ik}', while x_{ik}' is less than y_{ik}' for positive criteria. In order to simplify, $[min_{ik}', max_{ik}']$ is used to denote the scaled local constraint interval of $cons_k$ for task t_i. $Qmax(s_i.c_k, y_{ik})$ is the maximal value of constraint c_k among all values $s_{ij}.c_k$ in $\boldsymbol{SP}_i$ and y_{ik}, while $Qmin(s_i.c_k, x_{ik})$ is the minimal value among $s_{ij}.c_k$ and x_{ik}.

Definition 2. An upper bound service S_u is a virtual service whose QoS criteria values consist of a group of values max_{ik}'. For every task t_i, there exists a upper bound service S_{iu} , such that $\boldsymbol{C} = \{mak_{i1}', \ldots, max_{ik}', \ldots, max_{im}'\}$ on S_{iu}.

Definition 3. A lower bound servie S_l is a virtual service whose QoS criteria consist of the group of values min_{ik}'. For every task t_i, there exists a lower bound service S_{il}, such that $\boldsymbol{C} = \{min_{i1}', \ldots, min_{ik}', \ldots, min_{im}'\}$ on S_{il}.

For service s_{ij}, s_{ij} is filtered out of service spool $\boldsymbol{SP_i}$ to become qualified service for task t_i, iff $\forall k \in N^*, s_{ij}.c_k' \in [S_{il}.min_k', S_{iu}.max_k']$. Here, a qualified service refers to that both functional and non-functional properties of service s_{ij} are satisfied. In order to evaluate a qualified service s_{ij}, a utility function $Score(s_{ij})$, which converts multi-dimensional criteria into a sore in numeric form, is proposed as:

$$Score(s_{ij}) = \sum_{k=1}^{n} (s_{ij}.c_k' * w_{ik}). \tag{11}$$

The scores of S_{iu} and S_{il} are aslo calculated by formula (11) with a few modifications. Specifically, s_{ij} is substituted by S_{il} or S_{iu}, and $s_{ij}.c_k'$ is substituted by $S_{il}.min_{ik}'$ or $S_{iu}.max_{ik}'$ accordingly. w_{ik} refers to the weight of task t_k' criteria. It is because that the weights of criteria among different tasks are not the same generally. Take two services, bank service and shopping service, for example, reputation is the most important criterion for bank service in most scenarios, but price may play a more important role than others for shopping service. When customers fail to provide the weights for some reasons, it is more objective to reflect and model customers' most preferences through the weights of tasks and the weights of task's criteria. Finally, a qualified service s_{ij} with the highest score in the interval $[Score(S_{iu}), Score(S_{il})]$ is selected to implement the task t_i. The remaining tasks are fulfilled by repeating the same procedure.

4 A Case Study: Online Shopping

We consider an online shopping scenario (see Fig. 1) to validate our proposed method clearly. Fig. 1 provides a composite process composed of three tasks, namely, *T-shirt* service, *bank* service and *shipping* service. Assume that a customer needs a T-shirt with the requirements: the expected price range is [20$, 40$], the execution duration range is [0days, 7days] and the reputation scope is [2, 5] (the full range is [0, 5]). Thus service criteria set $\boldsymbol{C} = \{c_{price}, c_{duration}, c_{reputation}\}$. Firstly, a satisfactory T-shirt is ordered by *T-shirt* service. Secondly, a suitable *bank* service is chosen to deal with fees. Thirdly, a *shipping* service receives T-shirt and then transports T-shirt to customers. Next, our proposed method is applied to the scenario according to the following steps, each of which corresponds to the steps introduced in section 3.

1) Setp1: Task Weights Calculation

The weights of tasks are calculated based on formulae (1) (2). Assuming the tasks' weights has been calculated according to the first step of our method, with *T-shirt* service 0.7, *bank* service 0.1 and *shipping* service 0.2, namely $\boldsymbol{W} = \{0.7, 0.1, 0.2\}$.

2) Setp2: Global Constraints Decomposition

The local constraint of price for the first task *T-shirt* is calculated according to the following equations based on formulae (3), (4) and (5).

$$\begin{cases} x_{11} : x_{21} : x_{31} = 0.7 : 0.1 : 0.2 \\ x_{11} + x_{21} + x_{31} = 20 \end{cases} \qquad \begin{cases} y_{11} : y_{21} : y_{31} = 0.7 : 0.1 : 0.2 \\ y_{11} + y_{21} + y_{31} = 40 \end{cases}$$

The solution is x_{11}=14 and y_{11}=28. Therefore, the local constraint of price is [14$, 28$] for *T-shirt* service. In the same way, the local constraint of duration is calculated as [0days, 4.9days] and reputation is [1.4, 3.5] with the following equation based on formula (5).

$$F(x_{11},x_{21},x_{31})=\frac{1}{3}(x_{11}+x_{21}+x_{31})$$

3) Step3: Local Optimization Selection

The service with the high score in [*Score*(S_{1u}), *Score*(S_{1l})] is selected to implement *T-shirt* service. Suppose $\boldsymbol{SP}_1$ = {s_{11}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}}, the detailed specifications are shown in Table 1. All the values are scaled by *SAW* technique. Then the upper bound service S_{1u} and lower bound service S_{1l} can be identified by the scaled values of the local constraints. Afterwards, we calculate the scores of these services according to the formula (11) where the weights of QoS criteria of *T-shirt* service are provided by experts, with price 0.7, reputation 0.2 and duration 0.1, i.e.,W_1 = {0.7, 0.2, 0.1}. The results are shown in Table 2. s_{13}, s_{14} and s_{15} are filtered out of all the services to become qualified services, owing to each value of QoS criteria of them lies in the corresponding criterion interval between S_{1u} and S_{1l}. s_{15} with the highest score is selected to execute *T-shirt* service. the following equations:

$$\begin{cases} x_{21}:x_{31}=0.1:0.2 \\ x_{11}+x_{21}+x_{31}=20 \\ x_{11}=18 \end{cases} \qquad \begin{cases} y_{21}:y_{31}=0.1:0.2 \\ y_{11}+y_{21}+y_{31}=20 \\ y_{11}=18 \end{cases}$$

according to formulae (7), (8) and (5) are used to calculate the local constraint price of the second task *bank service*. Other local constraints are calculated in the same manner. Sequentially, the detailed selecting procedures are identical to that of the first service. As to the third service, it takes the same steps as the second service does. Finally, the services for executing the web service composition have been selected successfully, which as far as possible to maximize the probability of reaching customers' satisfaction.

5 Related Work and Comparison

Considerable methods to select services for web service compositions also have been widely discussed from many aspects based on technologies such as process[4], context [1], semantic [8] and so on. Zeng et al. [3] proposes local and global optimization approaches. The third step of our method is similar to the local optimization procedure of this paper, but we consider the virtual services upper/lower bound service and adopt the different weights which are task's criteria. Maamar et al. [9] propose a context model composed of *I*/*W*/*C*-Contexts. According to the model, the tasks' weights that we bring forward belong to *W*-Context. However, for our best knowledge, little attention has been paid to tasks' weights. With respect to service matching, Plebani et al. [10] provide an algorithm to evaluate the similarity degree between two web services from three aspects. These works [1][11] relating to service matching issues can be applied to our work for calculating the similarities of profiles between customers and tasks.

Table 1. Service pool ***SP_1***

Service	*Price*	*Duration*	*Reputation*
s_{11}	38	1	4
s_{12}	28	1	4
s_{13}	25	1	5
s_{14}	22	2	3
s_{15}	18	1	3
s_{16}	10	2	3

Table 2. Results for T-shirt Service

Service	*Pri*	*Rep*	*Dur*	*Score*
s_{13}	0.4643	0.5	0.7959	0.5046
s_{14}	0.5714	0.5	0.5918	0.55916
$\mathbf{s_{15}}$	**0.7143**	**0.25**	**0.5918**	**0.60919**
s_{1u}	0.8571	0.625	1	0.82497
s_{1l}	0.3571	0.1	0	0.26997

6 Conclusion

In this paper, we utilize two hierarchical weights, i.e., the weights of tasks and task's criteria, to partially reflect customers' preferences when they fail to give their preferences. The main feature of our method is that global constraints are decomposed successfully and reasonably into local constraints with the tasks' weights.

Acknowledgments. This paper is partly supported by the National Science Foundation of China under Grant No. 60721002, 60673017 and 60736015, and Jiangsu Provincial NSF Project under Grants BK2008017.

References

[1] Blake, B., Kahan, D., Nowlan, M.: Context-aware agents for user-oriented web services discovery and execution. Distributed and Parallel Databases 21, 39–58 (2007)

[2] Mohammad, A., Thomass, R.: Combining global optimization with local optimization for efficient QoS-aware service composition. In: 18th International Conference on World Wide Web (WWW), pp. 881–890 (2009)

[3] Zeng, L., Benatallah, B., Ngu, A.H.H., Dumas, M., Kalagnanam, J., Chang, H.: QoS-aware middleware for web services composition. IEEE Trans on Software Engineering 30(5), 311–327 (2004)

[4] Zeng, L., Ngu, A., Benatallah, B., Podorozhny, R., Lei, H.: Dynamic composition and optimization of web services. Distributed and Parallel Databases 24(1-3), 45–72 (2008)

[5] Cardoso, J., Miller, J., Sheth, A., Arnold, J.: Quality of service for workflows and web service processes. Journal of web semantics 1(3), 281–308 (2004)

[6] Segev, A., Toch, E.: Context-based matching and ranking of web services for composition. IEEE Transactions on Services Computing 2(3), 210–222 (2009)

[7] Yoon, K.P., Hwang, C.-L.: Multiple Attribute decision making: An introduction (Quantitative Applications in the Social Sciences). Sage Publications, Thousand Oaks (1995)

[8] Kim, I.W., Lee, K.H.: A model-driven approach for describing semantic web services: From UML to OWL-S. IEEE Trans. on Systems, Man, and Cybernetics 39 (2009)

[9] Maamar, Z., Mostef́aoui, S.K., Yahyaoui, H.: Toward an agent-based and context-oriented approach for web services composition. IEEE Trans. on Knowledge and Data Engineering 17(5), 686–697 (2005)

[10] Plebani, P., Pernici, B.: URBE: Web service retrieval based on similarity evaluation. IEEE Trans. on Knowledge and Data Engineering 21(11), 1629–1642 (2009)

[11] Medjahed, B., Atif, Y.: Context-based matching for web service composition. Distributed and Parallel Databases 21(1), 5–37 (2007)

Preference-Aware QoS Evaluation for Cloud Web Service Composition Based on Artificial Neural Networks

Xuyun Zhang[1,2] and Wanchun Dou[1,2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] Department of Computer Science and Technology, Nanjing University
210093, Nanjing, China
xyzhanggz@gmail.com, douwc@nju.edu.cn

Abstract. Since QoS properties play an increasingly important role during the procedure of web service composition in Cloud environment, they have obtained great interests in both research community and IT domain. Yet evaluating the comprehensive QoS values of composite services in accord with the consumers' preferences is still a significant but challenging problem due to the subjectivity of consumers. Since reflecting preference by the explicit weights assigned for each criterion is quite arduous, this paper proposes a global QoS-driven evaluation method based on artificial neural networks, aiming at facilitating the web service composition without preference weights. As well as this, a prototype composition system is developed to bolster the execution of proposed approach.

Keywords: Cloud; web services composition; QoS; preference; artificial neural networks**.**

1 Introduction

In recent years, the issues of cloud computing have been researched in depth by both the academia and IT domains. Cloud computing is an emerging computing paradigm, aiming at sharing resources that include infrastructure, software, application and business process [1]. As mentioned in [2], four types of resources can be accessed over the Internet: infrastructure, software, application and business process. Especially, the software application resources are provisioned and consumed via the Software as a Service (SaaS) [3] model. In cloud computing environment, although there are tremendous amount of web services available in the clouds, yet they have a limited utility when taken alone and might fail to accomplish the service users' multiple function requirements. For the purpose of satisfying the consumers' requirements and service resources reusing and sharing, it is essential to compose the single ones into a more powerful composite service.

It is essential to select the appropriate web services for users in terms of their preference, as the enormous amount of functional-equivalent web services exist in the clouds for certain task. QoS-aware composition is employed in this paper to guide the selection process. This issue has been investigated in depth and various QoS models and service selection algorithms are proposed [4], [5]. How to pick the suitable component web

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 410–417, 2010.

services to implement the composite one according to consumers' preference and maximize the consumers' content is a challenging problem.

Our work mainly aims at skirting the numeric preference weights directly in case no users' preference weights are provided, and then evaluating the candidate composite service relatively precisely. For the purpose that we are able to select the optimal component services for the composite schema, a preference-aware QoS evaluation approach is proposed based on the artificial neural networks (ANNs [6]). The approach consists of two phases. Firstly, ACP (Algebra of Communication Process [7]) expressions are employed to aggregate the value of each criterion of a composite service [8]. And then the artificial neural networks technology is introduced to simulate the weights for each criterion of the composite service. According to trained artificial neural network, the criteria values are combined into a comprehensive QoS value to evaluate the composite service.

The reminder of this article is organized as follows: The following section specifies background on the web service composition in Cloud environment and the artificial neural networks. Section 3 mainly proposes an algorithm to aggregate values of each criterion with ACP expressions, and the service composition is evaluated by combining the aggregated values into a one. In Section 4, a corresponding QoS-aware Cloud service composition prototype system is formulated. Section 5 introduces the related research work. Finally, Section 6 concludes this paper.

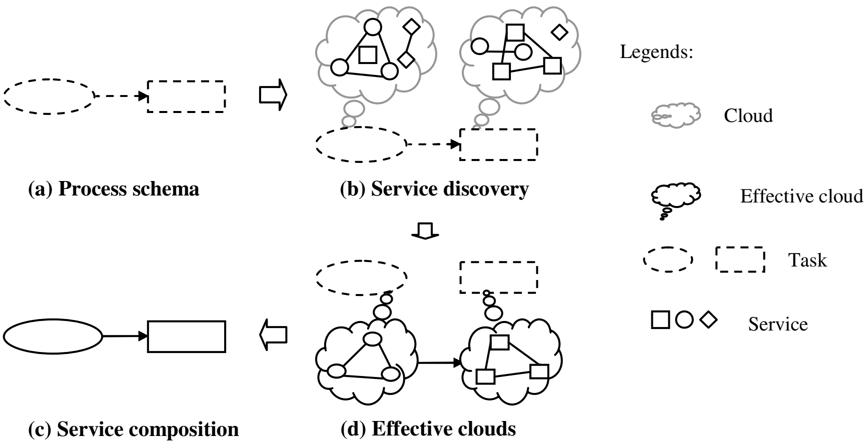
2 Background

2.1 Cloud Web Service Composition Scenario

As mentioned in the previous work [1], [2], [3], the models of services delivered in the Clouds could be mainly categorized into three types: Infrastructure as a Service (Iaas), Platform as a Service (PaaS), Software as a Service (SaaS). The web services referred to in this article are Cloud application services, i.e., we mainly focus on services delivered by SaaS model. Aiming at illustrating the web service composition process among clouds, Fig.1 demonstrates the composition scenarios in cloud surroundings step by step. To simplify the discussion, it is assumed that the web service composition schema is obtained either by analysts manually or by the AI planner automatically [9]. The process based composition of web services in Cloud environment could be elaborated as following definition.

Definition 1: (**Process based Cloud Web Service Composition**). In the Cloud environment, process based composition is a triple <Schema, Pools, Strategy>. Schema specifies the execution process of web service composition. Pools provide the candidate services to carry out the tasks in the predefined schema. Strategy refers to the techniques and mechanisms taken to implement the selection of the suitable web services for a task.

In this paper, schemas are described by ACP expressions. For briefness, please refer to literature [7], [8] if necessary. Then the qualified service selection problem (SSP) can be formulated as:



(a) Process schema

(b) Service discovery

(c) Service composition

(d) Effective clouds

Fig. 1. Web service composition scenarios among service clouds

(SSP) Find a candidate composite service *cs*:

$$\max Sf(cs) = \sum_{i=1}^{n} \lambda_i * q_i(cs)\,,\ cs = SF(s_1, s_2, \ldots, s_n)\ s_j \in SP_j,\ i, j \in N \tag{1}$$

s.t. $qf_i(cs) \leq (\geq)\ b_i,\ \sum_{k=1}^{n} \lambda_k = 1$

Apparently, determining the weights plays an imperative and significant role in SSP, since it concerns whether the combination of each criterion is consistent with the consumers' original preference.

2.2 ANNs

Artificial neural networks (ANNs) are non-linear dynamic network systems consisting of a lot of simple but highly distributed and parallel information processing units which are connected with the weights [6]. Multiple layer perceptrons (MLP) trained by back-propagation (BP) algorithm is one of the most widely used neural networks. This kind of neural network is capable of expressing complex mapping from input to output via training the weights by back-propagation algorithm. Thus this type of neural network is adopted in this paper.

Hecht-Nielson proved that any continuous function could be approximated with a three-layer perceptron [6]. This property makes this kind of neural network a powerful tool to solve the weight assigning problem. As stated before, the evaluation function *Sf*(*cs*) is in the SSP, then we can utilize BP neural network for approximating evaluation function *Sf*(*cs*), and then we can evaluate the candidate composite services. The neuron number of input layer and output layer will be identified by the specific application. Since it is supposed that there are n QoS criteria for a web service, the input neuron number should be n. Because the output of the neural network is the evaluation result, there should be one output neuron. Identifying the hidden neuron number is a complicated issue with no final verdict so far [10]. In terms of the research work of Charence N.W.Tan and Gerhard E.W. [11], the neural network with hidden layer neuron number equals the sum of input layer and output layer or with a pyramid-like architecture will work better. In our model, the hidden neuron number equals the input neuron number (n).

Let w_{ki} denote the weight between the neuron i in the input layer and neuron k in the hidden layer. w_{jk} denotes the weight between the k^{th} neuron in the hidden layer and the j^{th} neuron in the output layer. Here i, k = 1, 2, ..., n, j = 1. In this model, log-sigmoid function is adopted as the activation function, the log-sigmoid function is $f(x) = 1/(1+e^{-cx})$. The evaluation function is $Sf(Q) = WQ$, where Q is the input QoS criteria vector, and W is the weight vector, so the bias values of the neuron can be set as 0.

3 Preference-Aware QoS Evaluation with ANNs

3.1 Identifying Criteria Weights in ANNs

Now we employ the back-propagation algorithm [10] to train the neural network, the process of which includes forward propagation and back propagation. Mean square error $E[e^2]$ is introduced as the power function to measure the performance of the weight parameters. Then the power function can be expressed as $f(w) = E[(o\text{-}w^Tq)^2]$, where o is the desired output, w is a weight vector for a neuron in the hidden layer of the output layer, q is the input QoS criteria vector. Then LMS (Least Mean Squire error) algorithm (or δ-rule) is incorporated with the back-propagation algorithm to adjust the weights [10]. Generally, the weights are initialized by a set of small rand values. With the iterative process of adapting the weights, more and more preference information of the consumers is added to the neural network. And the weights finally keep steady, reflecting consumers' preference highly approximately.

Here we conclude the steps of the algorithm as follows.

1). Initialize the weights in the neural network with small random values.

2). Input a sample and calculate the output of each layer with following formula:

$$Y^l = sigmoid(Y^{l-1}W^l) \tag{2}$$

3). Compute the error of output layer:

$$E \leftarrow E + 0.5 * \| d\text{-}O \|^2 \tag{3}$$

4). Gain the local gradient:

$$\begin{cases} \delta_j = \frac{1}{2}[(d_j - O_j)sigmoid'(v_j)] \\ \delta_j = sigmoid'(v_k)\sum_k \delta_k w_k \end{cases} \tag{4}$$

5). Modify the weights via the following formula, where n is the current iteration.

$$w_{ij}(n) = w_{ij}(n\text{-}1)+\eta\, \delta_j\, q_j \tag{5}$$

6). If there are samples, countinue the steps 1)~5), else if error E is less than a given value, the whole alogrithm terminates. Otherwise, train the neural network with these samples which are re-permuted until error E is less than a preestablish threshold.

3.2 An Aggregation Algorithm for QoS Values Based on ACP

Since the web service composition discussed in our work is process based, it is essential to introduce a model tool to describe the schema. In this paper we take advantage of ACP expressions to represent the composition schema because of the algebraic properties of ACP. Based on the previous work in [8] we present an aggregation algorithm for QoS values. More details about ACP, please refer to [7], [8].

In order to calculate the aggregated values conveniently, the composition schema expression needs to be transformed into postfix notation (or Reverse Polish Notation (RPN)). Therefore, we concentrate more on aggregating QoS value via RPN. The procedure of aggregating the QoS values is formalized by the following algorithm.

```
Input: RS, QCS, V /* RS is the RPN of composition
       schema of the service composition. QCS is the QoS
       criteria set and V is the QoS value vector.*/
Output:  aggrValue array  /* a set of aggregated values
                              for the composite service*/
Stack; /* is used to store the temporary QoS values*/
for c in QCS
  while (RS ≠NULL) do
x←first symbol of RS;  /*scan the RS*/
switch (x)
      case service $s_i$:
          Push($v_{ic}$);        /* $v_{ic}$ ∈ V, value of c for $s_i$*/
      case operator o:   /* o∈ {•, +, ||}*/
          $q_1$←Pop();
          $q_2$←Pop();
          Push($q_1$ o′ $q_2$);   /* o′ ∈ {+, *, min, max}*/
 aggrValue[c] ←Pop();
 end for
```

3.3 QoS Evaluation with Users' Preferences

To reflect the preferences of users and facilitate users' input, it is sensible that users express their preference via giving an order of criteria, coupled with functional necessity and QoS constraints. Formally, we utilize vector: $< q_{i_1}, q_{i_2}, \cdots, q_{i_n} >$, where q denotes a QoS criterion and $q_{i_1} \geq q_{i_2} \geq \cdots \geq q_{i_n}$, to represent the preferences of users.

For sake of capturing the preferences of users more precisely, the users are categorized into different groups in view of their preference. As mentioned in [8], they can be divided into n! classes. Hence, for each user class, the structure and weights of an

ANN trained by the BP algorithm are stored for later use. It is facile to evaluate a candidate composite service with the neural network according to users' preferences.

With the evaluation values, it is feasible to compare the candidate composite services for a composition. The one who own higher value is adopted to implement the composition.

Now we conclude the steps to evaluate the QoS value for the web service composition. The steps are listed as follows.

Step1. Describe the execution process of the composition schema with the ACP expressions.
Step2. Train the neural networks with BP algorithm for each user category.
Step3. Aggregate each QoS value of a composite service with RPN of the composition schema by the aggregation algorithm.
Step4. Calculate the combined QoS value of all criteria with trained neural networks.
Step5. Select the optimal one with the highest QoS value.

4 A QoS-Aware Cloud Service Composition Prototype System

Considering that the web service evaluation process is not independent in a web service composition system, it is necessary to implement a prototype system integrating several modules as a whole.

For the purpose of demonstrating how the proposed approach works, a prototype system called ACP based Cloud Service Composition Prototype System (ab. ACSCP) is developed. The architecture of this prototype system is illustrated in Fig. 2.

This prototype system consists of the following core modules: User Interface (*UI*), Composition Engine (*CE*), Schema, Service Repository (*SR*), Selector, QoS Model (*QM*) and ANNs. *UI* interacts with the end users, for receiving the requirements and exhibiting the resultant composition. *CE* is the backbone of the system with the obligation to coordinate other components. Schema adopts ACP expression as the denotation of composition schema. SR contains a group of service pools. Service could be retrieved by accessing to SR. Selector works as the interface of web service selection

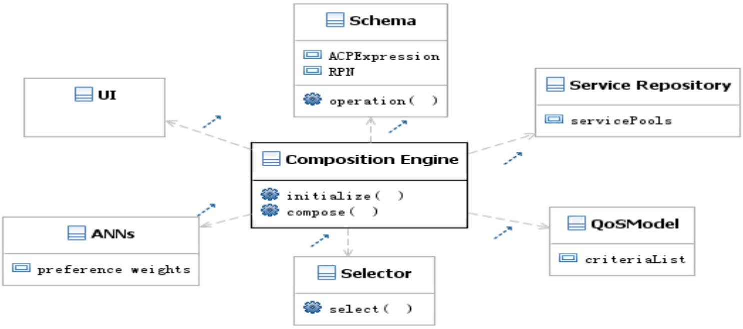


Fig. 2. Simplified UML class diagram for structure of ACSCP system

algorithms. *QM* stores the QoS information about services and provides other modules QoS-aware guidance. ANNs computes the preference weights of criteria and is responsible for adapting the weights dynamically.

The Composition Engine aggregates all other modules in Fig. 2, i.e., all other modules make up of the engine. Other modules interact with each other inside the engine, e.g., the algorithms in Selector module may access the web service parameter data in the Service Repository module. For concision, the association relationships between two modules are omitted.

5 Related Research Work

A multitude of previous work [4], [11] aimed at establishing QoS model and solving the QoS evaluation problem for web service selection. Yet they assumed that the weights of combining the QoS criteria are given either by users or experts, which could be boring and even infeasible in the real case. The work involved in [8], [12] attempted to acquire the numeric weights of users' preferences. In [8] the weights are gained according to the information of criteria order given by users, while the AHP approach is incorporated in [12] to gain the preference weights with the experts' scoring information.

To skirt the weights, G. Canfora et al [5] propose to adopt Genetic Algorithms to solve the QoS aware composition optimization problem. J. Chen et al [13] designed a new kind of neural network called decision neural network to assess the preferences of users. Similar to our work, L.Y. Guo et al [14] took advantage of artificial neural network to implement the QoS evaluation. However, since the users are not distinguished in terms of their preferences, the resultant ranking in their paper fluctuated and varied considerably. Due to the category of users in our work, this problem can be avoided. In [15] neural network is combined with genetic algorithm to solve the selection problem of composite web services.

6 Conclusion

In this paper, an ACP-based QoS value aggregation algorithm is represented for web service composition. To obviate tackling the explicit weights when the QoS criteria are combined, a neural network method is introduced to rate the QoS evaluation according to users' preference order information. The users are firstly categorized into various groups in terms of their preferences. And then neural networks are trained for each class. Additionally, a corresponding prototype system is implemented to support the web service composition in Cloud environment.

Acknowledgement

This paper is partly supported by the National Science Foundation of China under Grant No. 60721002, 60673017 and 60736015, and Jiangsu Provincial NSF Project under Grants BK2008017.

References

1. Zhang, L., Zhou, Q.: CCOA: Cloud Computing Open Architecture. In: IEEE International Conference on Web Services, pp. 607–616 (2009)
2. Zhang, L.J., Chang, C.K., Feig, E., Grossman, R., Panel, K.: Business Cloud: Bring the Power of SOA and Cloud Computing. In: IEEE International Conference on Service Computing (2008)
3. Armbrust, M., Fox, A., Griffith, R., Joseph, A.: R. Katz, Konwinski, Lee, A.G., Patterson, Rabkin, D.A., Stoica, I., Zaharia, M.: Above the Clouds: A Berkeley View of Cloud computing. Technical Report. University of California at Berkley, USA (2009)
4. Zeng, L., Benatallah, B., Ngu, A.H.H., Dumas, K.M.J., Chang, H.: QoS-Aware Middleware for Web Services Composition. IEEE Transaction Software Engineering 30(5) (2004)
5. Canfora, G., Penta, M.D., Esposito, R., Villani, M.L.: An Approach for QoS-aware Service Composition based on Genetic Algorithms. In: Proceedings of the Conference on Genetic and Evolutionary Computation, pp. 1069–1075 (2005)
6. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, pp. 181–200. Posts&Telecom Press, Beijing
7. Fokkink, W.: Introduction to Process Algebra. Springer, Heidelberg (2007)
8. Zhang, X.Y., Dou, W.C.: A Global QoS-Driven Evaluation Strategy for Web Services Composition. In: IEEE International Symposium on Parallel and Distributed Processing with Applications, pp. 595–600 (2009)
9. Peer, J.: Web service composition as AI planning - a survey. Technical report. University of St. Gallen, Switzerland (2005)
10. Nilsson, N.J.: Artificial Intelligence: A New Synthesis. Morgan Kaufmann Publishers, Inc., San Francisco (1998)
11. Charence, N.W.T., Gerhard, E.W.: QoS computation and Policing in Dynamic web service selection. In: Proceeding of the 13th International Conference on World Wide Web, pp. 66–73. ACM Press, New York (2004)
12. Lv, C., Dou, W.C., Chen, J.J.: QoS-Aware Service Selection using QDG for B2B collaboration. In: Proceeding of the 14th International Conference on Parallel and Distributed Systems. Melbourne, Australia (2008)
13. Chen, J., Lin, S.: A Neural Network Approach - Decision Neural Network (DNN) for Preference Assessment. IEEE Transaction on System, Man and Cybernetics (2004)
14. Guo, L.Y., Chen, H.P., Yang, G., Fei, R.Y.: A QoS Evaluation Algorithm for Web Service Ranking based on Artificial Neural Network. In: International Conference on Computer Science and Software Engineering (2008)
15. Yang, L., Dai, Y., Zhang, B., Gao, Y.: Dynamic Selection of Composite Web Services Based on a Genetic Algorithm Optimized New Structured Neural Network. In: Proceeding of the International Conference on Cyberworlds (2005)

Software Architecture Driven Configurability of Multi-tenant SaaS Application

Hua Wang and Zhijun Zheng

School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou, China
wanghua96@126.com, zjzheng9999@sina.com

Abstract. SaaS (Software as a Service) is a new emerging software application delivery model based on Internet. SaaS serves for multiple tenants with a list of business services to be delivered. The configurability of SaaS application has become an attractive aspect for tenants. The characteristics of the SaaS configurability have resulted in a recent drive to revisit the design of software architecture and challenges resulting from SaaS application. Existing approaches have made configurability strategies with external model that used formal method. The proposed method is novel because it utilizes the software architecture as a lever to coordinate between functional architectural elements and configurability components. By employing AOP (Aspect-oriented Programming), the method regards configurability as a crosscutting to realize configurability of SaaS application. Finally, a case study is assessed based on the proposed method.

Keywords: software architecture; configurability; multi-tenancy; SaaS.

1 Introduction

In traditional applications, configurability of system is required to realize flexible deployment in later software process. Configurability is driven by uncertain parameters in design process of software-intensive systems. Furthermore, ever-changing requirements, mutative runtime environments, diverse user preference, and different computing resources calls for configurability of relevant parameters to survive in software running context. This trend enables the managed system to be promoted according to given business goals under certain contexts so as to take the system adaptive to dynamic factors underlying in software evolution. The most valuable resource will be the limited resource of human attention [1]. How to minimize the consumption of the special resource and respond to the requirements of configurability more efficiently is a challenge.

SaaS (Software as a Service) is a new emerging software application service mode based on Internet. SaaS provides network infrastructure and operating platform of software and hardware for tenants. SaaS providers are in charge of application implement and maintain activity. SaaS customers need no more purchase software/hardware, build the machine room, and recruit new employees. SaaS is a kind of software layout model and accommodate facilities for customers to trusteeship, deployment and access using

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 418–424, 2010.

Internet. Compared to traditional software delivery, SaaS service can always be accessed where there is Internet. According to actual requirements, tenants lease software service. The SaaS delivery model essentially separates software ownership from the user—the owner is a vendor who hosts the software and lets the tenant execute it on-demand through some form of client-side architecture via the Internet or an Intranet[2].

SaaS serves for multiple tenants with a list of business services to be delivered. The key of SaaS software architecture with multi-tenant is to isolate data among of different tenants. The strategy guarantees that different tenants share the same running instance of SaaS application and that independent personal experience of application and data space are provided. The configurability of SaaS application can be implemented by software architecture driven model by using relevant architectural element---connector. In the context of multi-tenant, software architecture driven configurability mainly includes function configurability and user interface configurability.

The characteristics of the SaaS configurability have resulted in a recent drive to revisit the design of software architecture and challenges resulting from SaaS application. Software architecture provides a whole guideline to create target system, including specification of architectural elements, constraint of components and communication contract between architectural elements and software surviving environment. Existing approaches have made configurability strategies with external model that used formal methods, such as[3, 4]. The proposed method is novel because of it utilizes the software architecture as a lever to coordinate between functional architectural elements and configurability components. By employing AOP (Aspect-oriented Programming)[5], the method regards configurability as a crosscutting to realize configurability of SaaS application.

The remainder of this paper is organized as follows: Section II introduces the configurability of SaaS application. We discuss the function configurability and user interface configurability in our context. In Section III, configurability connector based on AOP is designed. In Section IV, we report on a case study on configurability method based on software architecture and Section V discusses conclusions and directions for future works.

2 Configurability of SaaS Application

There are a variety of different aspects about the configuration function between traditional product-based configuration and tenant-based configuration as illustrated in Table 1.

Table 1. Compare between traditional troduct-based and tenant-based configuration

Item	Product-based configuration	Tenant-based configuration
Relevant	Relevant to product features	Relevant to tenant features
Configuration Occasion	Before running system	At runtime of system
Configuration Quantity	One for the whole system	One for each tenant
Role of configuration	Service engineer	Tenant administrator
Scope	Valid for the whole system	Only valid for tenant data
Load time	Initial phase of system	Dynamic loading at runtime

The differences make it more difficult for software architect to design the configurability schemas. We discuss the function configurability and user interface configurability in our context as follows.

2.1 Function Configurability

The functions of traditional application are almost identical with user requirements. If the same application is delivered to diverse customers with significant requirement differences, the application modules will be recombined at deployment according to user requirements. However, with regard to SaaS application, it is almost impossible that the system functions of application are identical with each tenant. The majority of tenants demands partial function modules. It is unbelievable for each tenant to deploy a different version of application because all tenants use the same application. How can an SaaS application allows for different tenant to purchase different function modules set while satisfying online use?

SaaS application with multi-tenant emphasizes the idea of the need to use and the need to pay. The application supports the customers to purchase needed modules and different customers use different function modules set. In this way, function configurability of SaaS applications resolves the problem of requirement differences to be adaptive to diverse tenants crossing distinct business domains. For example, the domains of education, training, service and catering seldom emphasize on product management and consequently there is rarely order management. These domains think much of customer service and route tracking management. On the contrary, the domains of retail, distribution and manufacturing underline product and order management, while seldom emphasize customer service and route tracking management.

Consequently, SaaS applications need support function configurability and allows for customers to subscribe function modules. Online configuration at runtime enables different tenants operate diverse function at the same time. There are three steps for SaaS application to realize configurability as follows.

(1) Identify Atomic Function

The whole application should be divided into principal and independent atomic functions. All atomic functions compose all functions of the whole system. Function decomposition complies with the user value-driven principle, i.e., each atomic function provides certain value for users. If an atomic function gives no value to users, no user will purchase the function. For example, CRM application provides the creation function of customers including the check of an account and maintenance of contact. Both sub functions are only steps of creation of account without providing values to users. Consequently, both of them can not be defined as atomic functions.

After identifying all atomic functions, an important step is to define functions and dependency of functions because some functions relies on other functions. Take CRM application as an example again, the function of modification of orders of users is dependent on the function of query orders of users. If a customer do not purchase the latter function, he/she can not use the former function. Function definition means the description of the atomic function including the name, keywords, description and all dependency relationship. The relationship can be identified by function list. Component diagram of atomic function definition can designed as illustrated in Fig. 1.

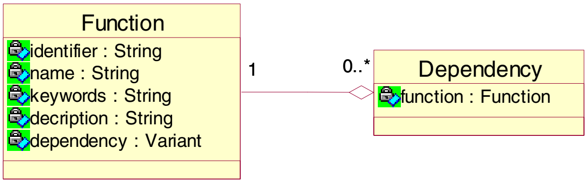


Fig. 1. Component diagram of atomic function definition

(2) Design function package

After all atomic functions are identified, it is impossible for tenants to perform any function if required because it is infeasible for tenants to configure these atomic functions to finish a certain routine task. An practicable option is to package atomic functions according to tenant type and usage scenario, and then choose an appropriate function package for tenants.

Function packages are designed in the light of tenant type and business logic. Atomic functions are assembled considering the tenant scenario and usage preference. An SaaS application is divided into hundreds of atomic functions resulting in the complicated management. There are interrelated atomic functions and operation of these atomic functions is not independent. Furthermore, atomic functions that are interdependent must be performed together.

The design of function package is similar to atomic function, that is to say, the name, keywords and description are defined including atomic function set and sub function package set, as illustrated in Fig.2.

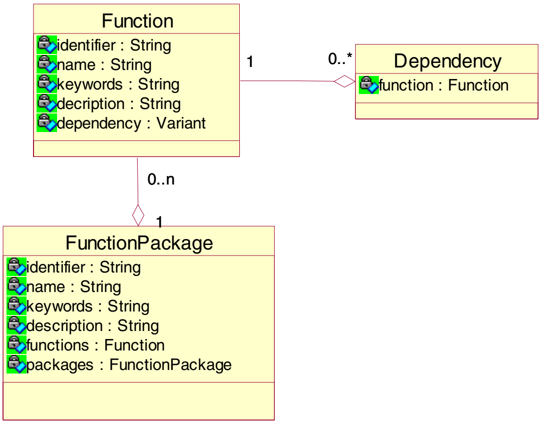


Fig. 2. Component diagram of function package definition

(3) Validation of function package

The validation process is necessary before deployment of function package. First, atomic function oriented validation process only be performed without considering the version of function package, which is ordered by tenants. Second, atomic functions are identified by tenants for usage and operation. This can be implemented by searching all function packages recursively according to SaaS application version that is ordered by tenants. Subsequently, all atomic functions included in the current function package can be identified. The validation process is regarded as NFR (non-functional requirement)[6]. NFR is considered as crosscutting concern (such as security, requirement and distribution) in AOP. The legal atomic function set ordered by tenants can be designed by AOP. The UI (User Interface) only displays functions which tenants ordered. The running environment of configurable systems decides user experience.

2.2 User Interface Configurability

The design of UI of traditional application almost satisfies the requirements of tenants. Even current UI does not meet their needs, developers can customize the UI according to the tenant requirements before deployment. However, under multi-tenant environment, if all tenants use the same graphic interface, many tenants do not approve of the default UI. How can a SaaS application support UI customization? The crux of the resolution is to realize the configurability of menus and content of pages.

(1) The configurability of menus

The function configurability discussed above suggests different tenants can purchase and operate different function set. Nevertheless, the names of menus with same functions could be different towards diverse tenants. For example, the menu name of customers management in CRM system may be renamed as the menu name of patients management in HIS (Hospital Information System)[7]. The dynamic configuration and display of menu name can be implemented by a menu configuration class diagram shown in Fig.3.

The hierarchical structure and layout of menus are diverse among different tenants. This calls for another configurability of system menus---a tenant has a set of menus; a menu associates with an atomic function; menus are organized as tree structure forming subordinate menu structures; and the show sequence of menus should be decided too.

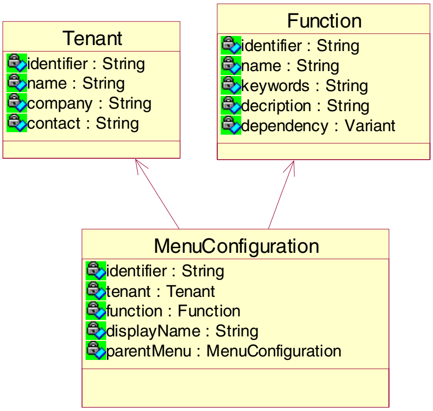


Fig. 3. Component diagram of function package definition

(2) The configurability of page elements

The page elements provide a interaction environment between tenants and SaaS application, which are similar to menus. Different tenants have divers requirements with regard to the numbers, location, sequence, layout and meaning of page elements. Take a CRM system as an example again, there is a label naming "customer name", while some tenants prefer to "agent name".

Extensible elements are displayed in pages by redefining these elements. Consequently, the number of page elements differs among different tenants. Elements assigned in design phase can not be deleted in most cases. But tenants permit some inessential page elements to be hidden.

3 Design Configurability Connector Based on AOP

As described above, configurability is a NFR and can be implemented by AOP. Software architecture is contaminated by the crosscutting concern of configurability resulting in the complexity of software architecture. By using AOP, configurability of SaaS application can be encapsulated as aspect. *Primary Component* performs some functional operation while *Aspect Component* performs configuration function. The Fig. 4 illustrates configurability model of SaaS application by *Aspect Component.*

In Fig.4, a configurability connector is constructed to isolate between *Function Component* and *Configuration Component*. At the same time , architectural elements are tied together by *Configurability Connectors* based on roles. Roles implies the

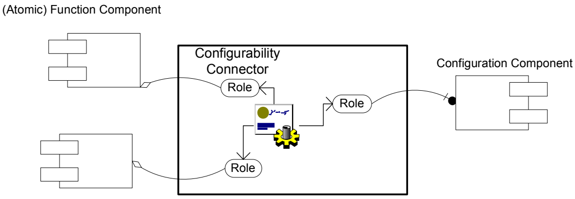

(Atomic) Function Component
Configurability Connector
Role
Role
Role
Configuration Component
(Atomic) Function Component

Fig. 4. Configurability connector model based on AOP

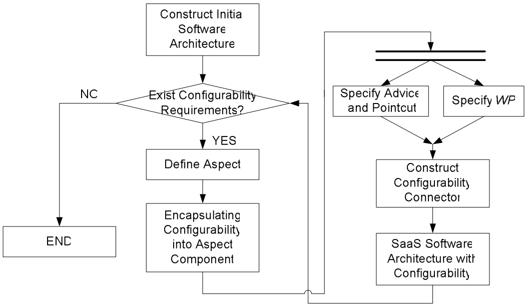

Construct Initia Software Architecture
NC
Exist Configurability Requirements?
YES
Define Aspect
Encapsulating Configurability into Aspect Component
END
Specify Advice and Pointcut
Specify WP
Construct Configurability Connector
SaaS Software Architecture with Configurability

Fig. 5. The construction process of Configurability Connector

activity of interaction between *Configurability Connector* and *Function Component*. The problem is how to construct a *Configurability Connector*. The key is to import the configurability of function and user interface. The configurability can be weaved into current software architecture by using Weaving Policy (*WP*) to instruct the dynamic behaviors from the initial SaaS application to be configurable one. *WP* is built by extensible *ECA* rules based on our previous works[8]. The construction process of Configurability Connector is shown in Fig. 5.

In Fig.5, *Advice* specifies when *Primary Component* is trigged to perform the operation of construction of configurability connector.

4 Conclusions and Future Works

Software architecture is employed as a lever to coordinate between functional architectural elements and configurability components. By employing AOP, the method regards configurability as a crosscutting to realize configurability and customization of SaaS application.

We believe more numerous test data should be obtained to ensure the proposed method enforces more effective performance. Additionally, the more aspects of configurability may be expanded by using more efficient interdisciplinary subject. Finally, we need to carry out more performance studies and conduct a more comprehensive evaluation to provide the best application service at the lowest cost, and this presents a challenge for future research.

Acknowledgments. This research is supported by the Scientific Research Foundation of Zhejiang University of Science and Technology under grant No. F501106902.

References

1. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. Computer 36, 41–50 (2003)
2. Laplante, P.A., Jia, Z., Voas, J.: What's in a Name? Distinguishing between SaaS and SOA. IT Professional 10, 46–50 (2008)
3. Wei, S., Xin, Z., Chang Jie, G., Pei, S., Hui, S.: Software as a Service: Configuration and Customization Perspectives. In: Congress on Services Part II, 2008. SERVICES-2, pp. 18–25. IEEE, Los Alamitos (2008)
4. Kwok, T., Thao, N., Linh, L.: A Software as a Service with Multi-tenancy Support for an Electronic Contract Management Application. In: IEEE International Conference on Services Computing, SCC 2008, pp. 179–186 (2008)
5. Colyer, A., Clement, A.: Aspect-oriented programming with AspectJ. IBM Systems Journal 44, 301–308 (2005)
6. Bagheri, H., Mirian-Hosseinabadi, S.H., Esfahani, H.C.: An Aspect Enhanced Method of NFR Modeling in Software Architecture. In: 10th International Conference on Information Technology (ICIT 2007), pp. 240–242 (2007)
7. Xudong, L., Huilong, D., Haomin, L., Chenhui, Z., Jiye, A.: The Architecture of Enterprise Hospital Information System. In: 27th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE-EMBS 2005, pp. 6957–6960 (2005)
8. Wang, H., Ying, J.: An Approach for Harmonizing Conflicting Policies in Multiple Self-Adaptive Modules. In: 2007 International Conference on Machine Learning and Cybernetics, pp. 2379–2384 (2007)

Parallel Accessing Massive NetCDF Data Based on MapReduce*

Hui Zhao[1,2], SiYun Ai[3], ZhenHua Lv[4], and Bo Li[1]

[1] Key Laboratory of Trustworthy Computing of Shanghai, China
[2] Institute of Software Engineering East China Normal University Shanghai, China
[3] School of EEE Communication Software & Network,
Nanyang Technology University Singapore
[4] Key Laboratory of Geographic Information Science, Ministry of Education,
Geography Department; East China Normal University, Shanghai, China
hzhao@sei.ecnuedu.cn, irisasy@sina.com
bli.sei.ecnu@gmail.com ,zhenhualv@gmail.com

Abstract. As a Network Common Data Format, NetCDF has been widely used in terrestrial, marine and atmospheric sciences. A new paralleling storage and access method for large scale NetCDF scientific data is implemented based on Hadoop. The retrieval method is implemented based on *MapReduce*. The Argo data is used to demonstrate our method. The performance is compared under a distributed environment based on PCs by using different data scale and different task numbers. The experiments result show that the parallel method can be used to store and access the large scale NetCDF efficiently.

Keywords: NetCDF, MapReduce, Data intensive , Parallel access.

1 Introduction

Science and engineering research has been becoming data-intensive work. How to use the data to analyze and improve human's living environment and protect disasters is what scientists devoted to[1]. How to manage massive scientific data becomes a challenge problem. NetCDF ,a National Science Foundation-sponsored program developed by Unidata has been widely used in terrestrial, marine and atmospheric sciences. The data described include single-point observations, time series, regularly-spaced grids, and satellite or radar images. Self-describing, good performance, portable and high usability are the main advantages [2-5]. The Climate and Forecast Metadata Conventions (CFMC) used by NetCDF has become part of NATO. As a unified data format, NetCDF will be used in more scientific applications in the near future [3]. With the increase of data, traditional NetCDF access method can't meet the need for managing

* This work is supported by the National Science Foundation of China under 60803077, Shanghai Leading Academic Discipline Project under B412 and Shanghai Pujiang Program under 08PJ1404600.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 425–431, 2010.

large scale scientific data efficiently. PNetCDF (Parallel NetCDF)[6] proposed by America Northwest University Argonne National Laboratory realizes parallel access for NetCDF based on MPI-I/O. This method encapsulates parallel communication mechanism and does little changes with initial functions. Compared with serial access, PNetCDF has greatly improved in performance. However, all the parallel communication needs to be controlled by programmer and users need to rewrite the code when they want to transfer their applications to different distributed system. MPI-based programming method also brings difficulties to users.

How to store TB scale NetCDF files reliably, how to locate the data rapidly and search the file in seconds, how to search a range of time climate data or search some climate data with a parameter rapidly have been become the key technologies in the fields of geosciences and climate. We implemented a new method MPNetCDF to store and access NetCDF data efficiently based on Hadoop[7-8] by using *MapReduce*[9] parallel programming model in a PC clusters. The experiment is realized by using Argo data for retrieving pressure information. The performance is compared by using different data scale on single PC and PC clusters. The results show that our method is efficient and applicable.

The rest of this paper is organized as follows. In section 2 we abstract the data structure of NetCDF. In section 3 we compare the performance for retrieving pressure information using Argo data in different number of computing nodes and data scale. Finally, we conclude our work and discuss future work.

2 NetCDF Data

NetCDF file itself is in the form of binary stream, can be accessed by API provided to access and transferred to text file in the form of CDL(Network Common Data Form Language). The structure of a NetCDF file presented by CDL can be understood easily .A NetCDF file has four parts: dimensions, variables, attributes and data. The function with multi arguments, $f(x, y, z) = value$ can be used to describe the NetCDF data structure. Arguments such as x, y, z are called dimension or axis. Value is called variables. Physical features and value of arguments like physics name, units, range, maximum and minimum value, default data are called attributes. We abstract the data model by the class diagram as showed in figure1.

NetCDF data model is like the tree structure which the root is group. Group can be looked as directory logically. Every dataset equals to a NetCDF file, which have three description types such as dimension, variable and attribute .Each of them is assigned a name and an ID. A NetCDF file has at least one group. In one dataset, there is at most one unlimited dimension. Variable is used for storing data, including normal variable and coordinate variable, whose type (dimension and attributes) must be declared before using. When a variable has no dimension, it is a scalar. A variable turns to be array-like structure and can store several values when the dimension is defined. Coordinate variables are a series of variables. They have the same name with their dimension. The definition of dimension only shows its length. There are two kinds of attributes that one

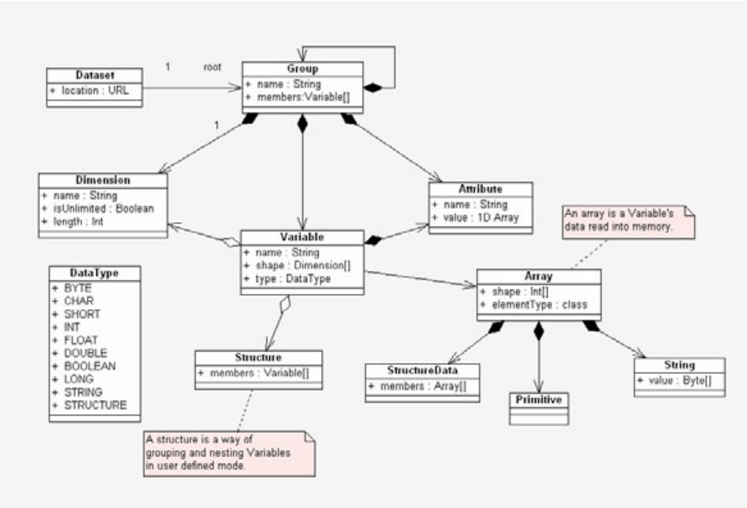


Fig. 1. NetCDF Class Diagram

is variable attribute which explains unit and range of variable, the other is global attribute which used for explaining the information of whole dataset. Global attribute doesn't need to add variable name in definition. The main part of a NetCDF file is data, which is stored in the form of one-dimension or two-dimension arrays by the sequence of variables' declaration.

3 Performance Analyses for NetCDF Retrieval with Argo Data

3.1 Parallel and Distributed Storage for NetCDF Data

First, we transfer NetCDF data into CDL file, then distribute large scale CDL files on HDFS(Hadoop Distributed File System). The System is composed of one Name-Node and several Data-Nodes. The CDL files are divided into size-fixed blocks(we chose 64MB) and distributed to store on the PC cluster in parallel. The system can locate the node that stores the data directly and in parallel according to the meta data structure recorded on Name-Node. We chose replication factor as 2 for reliable data storage, which guarantees data access when one node is in failure. Meanwhile name node uses heartbeat to periodicity get the response from each data node to check the working state of it. If any data node fails, name node will reschedule the computing task on the node ensuring good fault tolerance.

3.2 Parallel Access NetCDF Data

We use *IndexMapper* and *IndexReducer* to access NetCDF data parallel. I*ndexMapper* is in charge of parsing NetCDF file and mapping it into (*key, value*) pairs. There are four parameters: *key, value, OutputCollector* and *Reporter*. *Key* indicates current line's offset from initial line; *value's* type is an *NCDocument* class which is used for reading NetCDF files; *OutputCollector* sets the type of the intermediate (*key, value*) pairs and automatically collects all *Mapper* results; *Reporter* is responsible for reporting the running state of each *Mapper* task. *IndexReduce*r does reduce tasks, there are four parameters in the realization of reduce API: *key, values, OutputCollector* and *Reporter*. Types of key and values must be in accordance with the output type defined in *OutputCollector* of *Mapper* class. The values here are a collection having the same key. Through reading all intermediate data to do reduce job, it outputs the result by *OutputCollector*.

We define a *JobConf* class in *Main* function which is used for configuring parameter information of the system. Firstly, it passes one class to construction function of *JobConf*, and then Hadoop uses the class to find the jar file and distribute it to each node to run these tasks. Furthermore, it sets the *inputFormat*, input path and output path. Then the class that run *Map* and *Reduce* tasks is set. Finally the job is deployed to run on Hadoop.

3.3 Experiment Environment and Experiment Data

We do the experiments on the cluster environment which composed of one master node and eight data nodes. Master node is configured as follows: 2*4 CPU, 2.0GHz/CPU, 4GB RAM. Each data nodes is configured as follows: 2*4 CPU, 2.0GHz/CPU, 8GB RAM, 20TB storage capacity using NAS-on-SAN architecture with RAID 1+0. The master node runs on windows server 2003 and uses Red Hat Linux AS 4 as virtual machine. All nodes install Hadoop0.19.1 and JDK 1.6.0.6 as running environment.

We uses Argo (Array for Real-time Geostrophic Oceanographic) data[10] as experiment data to retrieval pressure value. Argo is on behalf of the observation network of ocean, taking the important tasks of global ocean observation.

3.4 Retrieve Pressure Data under Different Running Environment

We use "*PRES_ADJUSTED*" as keyword on data sets with different scale. The retrieval time is showed as figure 2.

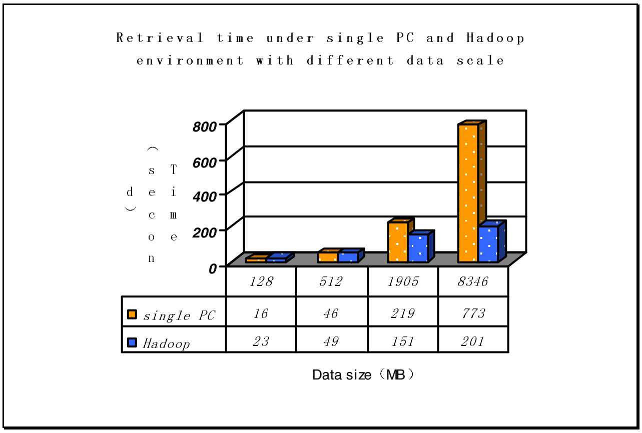


Fig. 2. The retrieval time under different run environment

With the increase of data scale, the method we proposed can save more time. It is more suitable for large scale data information processing. Otherwise, the method is not suit for small size data that the performance is worse than that of single node. The reason is that our method shows its distributed parallel advantages when dealing with large scale data (generally beyond GB or TB scale). When data scale is small, the retrieval performance declines for the reason that the system running on HDFS needs to consume the time for distributing tasks and get data's location information through different nodes to access data

3.5 Retrieve Pressure Data under Different Parallel Task Numbers

We also use the same keyword *"PRES_ADJUSTED"* to search in cluster environment. We use the dataset about 1.86G, 8.15G and 31.1G under 8, 6, 4, 2 data nodes respectively. We adjust the number of *Map* and *Reduce* tasks to improve the whole performance while number of data nodes is determined. We find that with the increase of data scale, raising the number of reduce tasks properly can improve the speed of retrieval when the number of map tasks is under the control of *InputFormat*. Figure 3 shows that retrieve time under different number of data nodes and figure 4 shows the retrieval time with 6 data nodes under different task numbers.

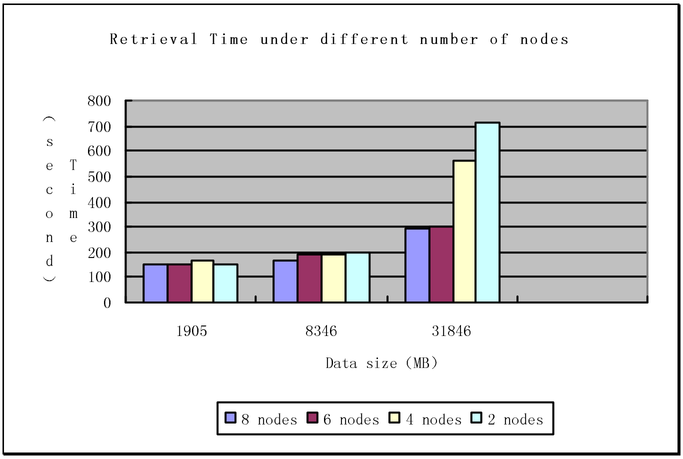


Fig. 3. The retrieval time under different node numbers

With the increase of data nodes, the retrieve time consumed decreases when data size doesn't change. When number of map tasks is determined, the whole system performance can be improved by adding reduce tasks properly. Such phenomenon is more obvious when data scale increases. The reason is that the number of map tasks has been determined by *InputSplit* of *InputFormat*, which can't change through outer parameter. However, we can control number of reduce tasks to improve system performance. It is more flexible to reallocate jobs for the failed nodes to improve load balance effectively in Hadoop environment.

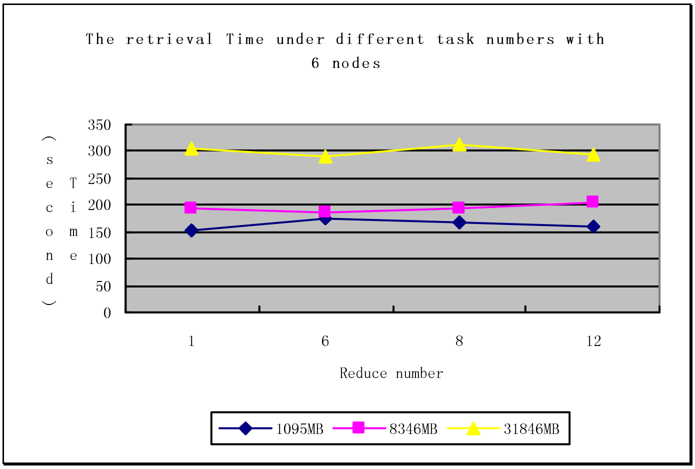


Fig. 4. The retrieval time under different task numbers with 6 nodes

4 Conclusions

Climate, ocean and satellite remote sensing information play more and more important role in analyzing and improving human's living environment and protecting disasters as well. NetCDF has been widely used in terrestrial, marine and atmospheric sciences with advantages like self-describing, good performance, portable and high usability. We think NetCDF will be applied in more fields as a unified data format in the near future. With the rapid increase of data scale, parallel accessing NetCDF become the urgent demand. We implemented a parallel method for storing and accessing NetCDF data This method based on *MapReduce* .The performance is analyzed using massive Argo data under a distributed environment based on PCs cluster installed Hadoop. The access performance under the circumstances of different data scale and different parallel task numbers is compared and analyzed. The experiments show our method is feasible, effective and efficient. Compared with other parallel programming model like MPI, *MapReduce* paradigm deals with user's parallel process automatically just by two functions: Map and Reduce. The fault tolerance and load balance mechanisms are provided. Furthermore, compared with MPI, programmers will have fewer difficulties to parallel their applications using *MapReduce*. So the programmer can devote to the realization of the business logic. As the Hadoop becoming mature, the method we proposed will be a feasible routine to manage massive NetCDF data. For the further work, we will research the method to access raw NetCDF files directly.

References

1. Gra, J., Liu, D.T., Nieto-Santisteban, M., Szalay, A., DeWitt, D.J., Heber, G.: Scientific Data Management in the Coming Decade. SIGMOD Record 34(4) (December 2005)
2. Rew, R., Davis, G., Emmerson, S., Davies, H., Hartnett, E.: The NetCDF Users' Guide Version 4.0, `http://www.unidata.ucar.edu/software/netcdf/docs/`

3. Zender, C.S.: Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). Environmental Modelling & Software 23, 1338–1342 (2008)
4. Jie, L.: Research and Development on the Key Technology of Marine Data Sharing Platform. Tianjin University (2008)
5. Jun-tao, J., Chan-yuan, M., Hai-ying, S., Qiang, L., Ji-chuan, T.: Setting up and Managing of Seafloor Terrain Grid Model Based on NetCDF. Hydrographic Surveying and Charting 27(5) (2007)
6. Li, J., Liao, W., Choudhary, A.: Parallel netCDF: A High Performance Scientific I/O Interface. ACM, 479–492 (2003)
7. The Hadoop Project, `http://hadoop.apache.org/`
8. Venner, J.: Pro Hadoop. United States of America, Apress (2009)
9. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM 51(1), 107–113 (2008)
10. China Argo Real-time Data Center, `http://www.argo-cndc.org/argo-china/legend/ARGO%20DATA.htm`

On Well-Formedness Rules for UML Use Case Diagram

Noraini Ibrahim[1], Rosziati Ibrahim[1], Mohd Zainuri Saringat[1],
Dzahar Mansor[2], and Tutut Herawan[3]

[1] Department of Software Engineering
Faculty of Information Technology and Multimedia
Universiti Tun Hussein Onn Malaysia
[2] Microsoft Malaysia
[3] Department of Mathematics Eduction
Universitas Ahmad Dahlan, Yogyakarta 55166, Indonesia
{noraini,rosziati,zainuri}@uthm.edu.my,
dmansor@microsoft.com, tutut81@uad.ac.id

Abstract. A software model is a widely used technique to specify software. A UML model may contain different diagrams and a diagram is built from different elements. Each element is restraint to certain constraint or well-formedness rules (WFR). Assurance to these WFR is important to ensure the quality of UML diagrams produced. Even though, the formal definition to UML elements is rapidly increased; there is still lack of formalization of WFR. Therefore, this paper will define the WFR for use case diagrams as it is ranked as one of the most used diagram among UML practitioners. The formalization is based on set theory by logic and quantification. Based on an example of a use case diagram, we show how the diagram satisfied the WFR. Then, the elements involved in the well-formedness problem are detected and formally reasoned.

Keywords: UML; Well-formedness rules; Use case diagram; Set Theory.

1 Introduction

A software model is an abstract and graphical representation of software functionalities and constraints. A model may consists of different diagrams [1]. In object oriented based system, requirements of the software is visualized, captured and documented using Unified Modeling Language (UML). Currently, UML is represented by 13 (thirteen) diagrams. UML use case diagram is one of the most used diagrams among UML practitioners [2]. It is made up of multi elements. Therefore, precise meaning of the elements is very important in order to have a common understanding of their meaning.

There are many researchers involved in giving formal definition to UML use case diagrams. Shinkawa [3] gives definition of use case, actor and association of them using Colored Petri Nets. Liu *et al.* [4] shows the formalization of use case model in terms of dynamic semantic. Overgaard and Palmkvist [5] also formalize the dynamic semantics of use cases and their relationships using operational semantics. Chanda *et al.* [6] defines elements of UML use case diagram using Context Free Grammar.

F.L. Wang et al. (Eds.): WISM 2010, LNCS 6318, pp. 432–439, 2010.

While Mostafa *et al.* [7] and Sengupta *et al.* [8] give formal definition to use case diagram using Z specification language.

However, there is still lack of formal definition of well-formedness rules (WFR) of the previous methods [3,4,5,6,7,8]. WFR are restriction of each element of UML diagrams to some constraints. It is important to satisfy the WFR within a single UML diagram before we proceed to other validation activities such as detecting and handling inconsistencies between diagrams as they impact the completeness of UML model [9]. In UML standard [10], WFR are described as natural language and some of them are specified using Object Constraint Language (OCL). Not all WFR in the standard are specified in OCL because it is too limited to express [11].

Therefore, in this paper, the WFR using set theory by logic and quantification will be defined. It is based on an example of a use case diagram and we will show how the diagram fulfilled the WFR. Furthermore, the elements involved in the well-formedness problem are detected and formally reasoned.

The rest of this paper is organized as follows. The discussion on related works is in Section 2. Section 3 presents the well-formedness rules of use case diagram, Section 4 on formalization of well-formedness rules of use case diagram using proposed technique and an example of use case diagram and elements involved in well-formedness problem, and Section 5 concludes the paper.

2 Preliminaries

This section is divided into two sub-sections. The first section is on formal definition of UML use case diagram and second section on WFR.

2.1 Formal Definition of UML Use Case Diagram

Even though there are various researchers involved in giving formal definition to UML use case diagrams, they focus to several domains. Elements of use case diagram such as actor and use case are specified as place and transition in CPN[3]. They are then used to check consistency with other UML diagrams without checking the well-formedness of use case diagram. While Liu *et. al.* [4] define functional semantics of use case diagram using operational semantics. Then, based on the formal definitions, the researchers check requirements consistency between use case diagram and class diagram without checking the fitness of the use case diagram to its constraints. Semantics of use case, uses and extends are formalized using an object-oriented specification language named ODAL [5]. They define the elements in terms of operations and methods instead of restraint of each element towards its constraint. Use case, actor and the relationship between them are expressed as CFG [6]. They describe one of the rules as actor has relationship with actor, which is either `<<include>>` or `<<extend>>`. This rule disagree with WFR as in UML standard [10] whereby `<<include>>` or `<<extend>>` involved binary relationship between use cases. Mostafa *et al.* [7] provide formal specification for elements of use case diagram using Z specification language. However, they do not formalize WFR that states generalization of use cases and actors are irreflexive. The Z specification also used by [8] to formalize use case, event and use case relationship. But they do not detail out the elements in terms of well-formedness as they focus on dynamic semantic of use case diagram.

2.2 Well-Formedness Rules

WFR or syntactical constraint of UML model is very important to ensure the quality of UML model [9,12]. There are many perspectives on well-formedness rules. Some define it as restriction on a single diagram [9] and there is researcher define it as consistency rules which is between different diagrams [13]. While in UML standard [10] specified WFR as constraints for elements of UML diagrams. This paper refers to WFR as in the standard. WFR are described as natural language and some of them are supported by Object Constraint Language (OCL). There are not all elements of UML diagrams has WFR. In scope of UML use case diagram, elements like <<include>> do not have WFR and certain WFR do not supportive by OCL since it is too limited to express [11].

3 The Proposed Technique

3.1 Well-Formedness Rules of Use Case Diagram

In this section, we describe Well-Formedness Rules of Use Case Diagram.

3.1.1 Actor

- An actor must have a name.
- Each actor must be associated/ involved with at least one use case.
- Actors are not allowed to interact (associate) with other actors.

3.1.2 Use Case

- A use case must have a name.
- Every use case is involved with at least one actor.

3.1.3 Association

- Association can only happen between actors and use cases.

3.1.4 Generalization

- Generalization can only happen between actors or between use cases.
- It is an acyclic, irreflexive and transitive relationship.

3.1.5 <<include>> Relationship

- <<include>> are relationship between use cases that have source use case and destination use case.
- Source use case is including use case and destination is included use case.

3.1.6 <<extend>> Relationship

- `<<extend>>` are relationship between use cases that have source use case and destination use case.
- Source use case is extending use case and destination is extended use case.

3.2 The Proposed Technique

This section summarizes the formal syntax of modeling notations of use case diagram. Before we describe the proposed technique, we will recall the fundamental concepts of relation in set theory. Let *S* be a non empty set and $A, B \subseteq S$. A *Cartesian product* of *A* and *B* denoted by $A \times B$ is defined by $A \times B = \{(a,b) | a \in A, b \in B\}$. A *binary relation S* from a set *A* to a set *B* is a subset of the Cartesian product $A \times B$. If $(a,b) \in S$, we write aSb and say that *a* is related to *b*. A relation *S* on *A* is said to be *reflexive* if $aSa, \ \forall a \in A$. Meanwhile, a relation *S* on a set *A* is said to be *irreflexive* if $a\$a, \ \forall a \in A$. A relation *S* from *A* to *B* is said to be *symmetric* if $aSb,$ then $bSa, \ \forall a,b \in A$. Meanwhile, a relation *S* on a set *A* is *asymmetric* if whenever $aSb,$ then $b\$a, \ \forall a,b \in A$. A relation *S* on *A* is said to be *transitive* if aSb and bSc , then $aSc, \ \forall a,b,c \in A$.

3.2.1 Formalization on Well-Formedness Rules of Use Case Diagram

This section summarizes the formal syntax of modeling notations of use case diagram.

Definition 1. *A use case diagram, U consists of four elements A, C, S and G where* $A = \{a_1, a_2, a_3, \cdots, a_m\}$ *is finite set of actors,* $C = \{c_1, c_2, c_3, \cdots, c_m\}$ *is finite set of use cases,* $S = \{s_1, s_2, s_3, \cdots, s_m\}$ *is finite set of associations and G is generalization.*

Note that every rule involves actor and use case. The property of association between actor and use case is given in the following property.

Property 1. *An actor and a use case must have a name.*

Property 2. *A use case diagram must have at least an actor and a use case.*

Property 3. *Let a be an actor, each actor must be associated (involved) with at least one use case and every use case is involved with at least one actor, i.e.*

$$\forall a \in A, \exists c \in C, \ and \ s(a) = c \ . \qquad (1)$$

Property 4. *Let* $A = \{a_1, a_2, a_3, \cdots, a_m\}$ *be a finite set of actors. Actors are not allowed to interact (associate) with other actors.*

$$\forall a_i, a_j \in A, \ s(a_i) \neq a_j \ , for \ 1 \leq i, j \leq m \qquad (2)$$

Property 5. *Association can only happen between actors and use cases. Thus conforming*

$$S \subseteq A \times C \tag{3}$$

Definition 2. *Acyclic is defined as a asymmetric relation in a generalization, i.e. among use cases and among actors.*

Definition 3. (Generalization between actors). *A generalization between two actors in $A = \{a_1, a_2, a_3, \cdots, a_m\}$ is acyclic, i.e. G is a relation on A $(G \subseteq A \times A)$, where aGb if and only if G is irreflexive, asymmetric and transitive, $\forall a, b \in A$.*

Note that, if in G, $a_i G a_j$, for $i \neq j$, then $a_j \not{G} a_i$. For instance, if we let $A = \{a_1, a_2, a_3\}$, then we have $G = \{(a_1, a_2), (a_1, a_3), (a_2, a_3)\}$.

Definition 4. (Generalization between use cases). *A generalization between two use cases in $C = \{c_1, c_2, c_3, \cdots, c_m\}$ is acyclic , i.e. G is a relation on C $(G \subseteq C \times C)$, where cGd if and only if G is irreflexive, asymmetric and transitive, $\forall c, d \in A$.*

Note that, if in G, $a_k G a_l$, for $k \neq l$, then $a_l \not{G} a_k$. For instance, if we let $B = \{b_1, b_2, b_3\}$, then we have $G = \{(b_1, b_2), (b_1, b_3), (b_2, b_3)\}$.

Include Relationship

Definition 5. *Let $C = \{c_1, c_2, c_3, \cdots, c_m\}$ be a finite set of use cases. A* <<include>> *relationship denoted by $\rightarrow$ is identified by $\rightarrow C \times C$, as $c_m \rightarrow c_n$, where c_m is source (including) use case and c_n is destination (included) use case.*

Extend Relationship

Definition 6. *Let $C = \{c_1, c_2, c_3, \cdots, c_m\}$ be a finite set of use cases. A* <<extend>> *relationship denoted by $\leftarrow$ is identified by $\leftarrow C \times C$, as $c_k \leftarrow c_l$, where c_k is source (extending) use case and c_l is destination (extended) use case.*

Definition 7. *Let $C = \{c_1, c_2, c_3, \cdots, c_m\}$ be a finite set of use cases. The extended use case c_l is defined independently of the extending use case c_k. Furthermore, c_l is meaningful independently of the extending use case c_k. The extending use case c_k typically defines behavior that may not necessarily be meaningful by itself.*

4 The Result

This section shows an example of use case diagram and how the diagram satisfies WFR. If the diagram is not well-formed, the violated WFR will be referred and the element involved in the violated is described.

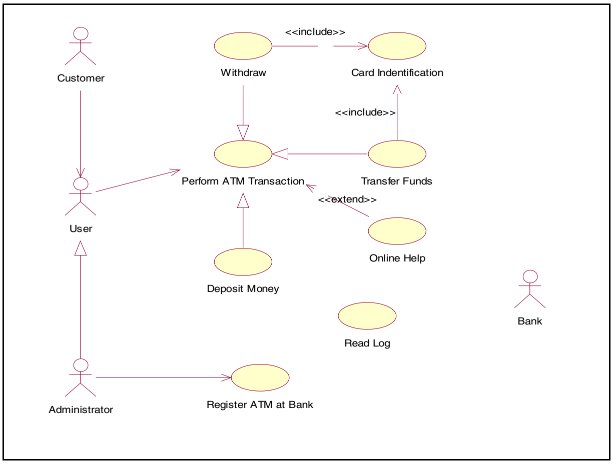


Fig. 1. Use Case Diagram of ATM System

Based on Fig. 1,

a. Once Actor Administrator is generalized to actor User, actor User cannot be generalized to actor Administrator. This fulfilled WFR that say generalization is asymmetric, i.e.

$$\text{Administrator}\, G\, \text{User},\ \ \text{but}\ \ \text{User}\, \not G\, \text{Administrator}\,.$$

b. An actor Administrator, Customer and User cannot be generalized to itself. This fulfilled WFR that say generalization is irreflexive, i.e.

$$\text{Administrator}\ \not G\, \text{Administrator}\,, \text{Customer}\ \not G\, \text{Customer}\quad \text{and}$$

$$\text{User}\ \not G\, \text{User}\,.$$

c. Actor User associate to Perform ATM Transaction use case. This fulfilled WFR that say an actor must associate to at least one use case.

$$\text{User}\ \ G\, \text{Perform ATM Transaction}\ .$$

d. Even though actor Customer not associated to any use case (seems it violates WFR that say an actor must associate to at least one use case), actor Customer is generalized to actor User, so actor Customer also associated to Perform ATM Transaction use case. Generalization is transitive.

$$\text{Customer}\ \ G\, \text{User and User}\, G\, \text{Perform ATM Transaction}\ \ ,$$

then

$$\text{Customer}\ \ G\, \text{Perform ATM Transaction}\ .$$

e. Actor Administrator associated to Register ATM at Bank. This fulfilled WFR that say an actor must associate to at least one use case.

$$\exists \text{Register ATM at Bank},\quad \text{Administrator}\, G\, \text{Register ATM at Bank}\,.$$

f. Perform ATM Transaction use cases cannot generalize to Withdraw, Transfer Funds and Deposit Money. This fulfilled WFR that say generalization is asymmetric, i.e.,

$$\text{Perform ATM Transaction } G \text{ Withdraw , but}$$
$$\text{Withdraw } \not{G} \text{ Perform ATM Transaction .}$$
$$\text{Perform ATM Transaction } G \text{ Transfer Fund , but}$$
$$\text{Transfer Fund } \not{G} \text{ Perform ATM Transaction .}$$

and

$$\text{Perform ATM Transaction } G \text{ Deposit Money , but}$$
$$\text{Deposit Money } \not{G} \text{ Perform ATM Transaction .}$$

5 Conclusion

UML is a popular modeling technique especially in object-oriented based software development. Existing UML formalization is focuses on dynamic semantics. Even though WFR is important to ensure the quality of UML diagram, there is still lack of research on giving formal definition of WFR. It is important to formalize the UML well-formedness rules in order proof the elements involved in the well-formedness problem. With this motivation, we have been described the UML well-formedness rules for use case diagram and formalize them using set theory. We intend to formalize WFR of other UML diagrams as the platform to conduct other validation activities such as detecting and handling inconsistencies between diagrams.

Acknowledgement

The authors would like to thank Ministry of Science, Technology and Innovation Malaysia for supporting this research under the ScienceFund Grant.

References

1. Huzar, Z., Kuzniarz, L., Reggio, G., Sourrouille, J.L.: Consistency Problems in UML-Based Software Development UML Modeling Languages and Applications. In: Jardim Nunes, N., et al. (eds.) UML Satellite Activities 2004. LNCS, vol. 3297, pp. 1–12. Springer, Heidelberg (2005)
2. Dobing, B., Parsons, J.: Dimensions of UML Diagram Use: A Survey of Practitioners. Journal of Database Management 19(1), 1–18 (2008)
3. Shinkawa, Y.: Inter-Model Consistency in UML Based on CPN Formalism. In: 13th Asia Pacific Software Engineering Conference (APSEC 2006), pp. 414–418. IEEE Press, Los Alamitos (2006)
4. Li, X., Liu, Z., He, J.: Formal and use-case driven requirement analysis in UML. In: 25th Annual International Computer Software and Applications Conference (COMPSAC 2001), pp. 215–224. IEEE Press, Los Alamitos (2001)

5. Övergaard, G., Palmkvist, K.: A Formal Approach to Use Cases and Their Relationships. In: Bézivin, J., Muller, P.-A. (eds.) UML 1998. LNCS, vol. 1618, pp. 406–418. Springer, Heidelberg (1999)
6. Chanda, J., Kanjilal, A., Sengupta, S., Bhattacharya, S.: Traceability of Requirements and Consistency Verification of UML Use Case, Activity and Class diagram: A Formal Approach. In: International Conference on Methods and Models in Computer Science 2009 (ICM2CS 2009), pp. 1–4. IEEE Press, Los Alamitos (2009)
7. Mostafa, A.M., Ismail, M.A., El-Bolok, H., Saad, E.M.: Toward a Formalization of UML 2.0 Metamodel using Z Specifications. In: Eighth ACIS International Conference on Artificial Intelligence, Networking, and Parallel/Distributed Computing, SNPD 2007, pp. 694–701. IEEE Press, Los Alamitos (2007)
8. Sengupta, S., Bhattacharya, S.: Formalization of UML Diagrams and Their Consistency Verification- A Z Notation Based Approach. In: Proceedings of the 1st India Software Engineering Conference, pp. 151–152. ACM, New York (2008)
9. Lange, C., Chaudron, M.: An empirical assessment of completeness in UML designs. In: 8th International Conference on Empirical Assessment in Software Engineering (EASE 2004), pp. 111–119. IEEE Seminar Digests (2004)
10. Object Management Group (OMG): OMG Unified Modeling Language^{TM} (OMG UML) Superstructure. Object Management Group (2009)
11. Lucas, F.J., Molina, F., Toval, A.: A Systematic Review of UML Model Consistency Management. Information and Software Technology 51(12), 1631–1645 (2009)
12. Lange, C.F.J., Chaudron, M.R.V.: Managing Model Quality in UML-based Software Development. In: 13th IEEE International Workshop Software Technology and Engineering Practice, pp. 7–16. IEEE Press, Los Alamitos (2005)
13. Labiche, Y.: The UML Is More Than Boxes and Lines. In: Chaudron, M.R.V. (ed.) Models in Software Engineering. LNCS, vol. 5421, pp. 375–386. Springer, Heidelberg (2009)

Author Index