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

LNCS 6988

Web Information Systems and Mining

International Conference, WISM 2011
Taiyuan, China, September 2011
Proceedings, Part II

2 Part II

Springer

Lecture Notes in Computer Science 6988

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

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

Web Information Systems and Mining

International Conference, WISM 2011
Taiyuan, China, September 24-25, 2011
Proceedings, Part II

Springer

Volume Editors

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

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

Junjie Chen
Taiyuan University of Technology, School of Computer and Software
Taiyuan 030024, China
E-mail: chenjj@tyut.edu.cn

Jingsheng Lei
Shanghai University of Electric Power
School of Computer and Information Engineering
Shanghai 200090, China
E-mail: jshlei@126.com

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

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-23981-6 e-ISBN 978-3-642-23982-3
DOI 10.1007/978-3-642-23982-3
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011936228

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

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

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

Printed on acid-free paper

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

Preface

The 2011 International Conference on Web Information Systems and Mining (WISM 2011) was held during September 24–25, 2011 in Taiyuan, China. WISM 2011 received 472 submissions from 20 countries and regions. After rigorous reviews, 112 high-quality papers were selected for publication in the WISM 2011 proceedings. The acceptance rate was 23%.

The aim of WISM 2011 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 collaborations. 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 Taiyuan University of Technology 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, 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.

September 2011

Gong Zhiguo
Xiangfeng Luo
Junjie Chen
Jingsheng Lei
Fu Lee Wang

Organization

Organizing Committee

General Co-chairs

Wendong Zhang	Taiyuan University of Technology, China
Qing Li	City University of Hong Kong, Hong Kong

Program Committee

Co-chairs

Gong Zhiguo	University of Macau, Macau
Xiangfeng Luo	Shanghai University, China
Junjie Chen	Taiyuan University of Technology, China

Steering Committee Chair

Jingsheng Lei	Shanghai University of Electric Power, China

Local Arrangements Co-chairs

Fu Duan	Taiyuan University of Technology, China
Dengao Li	Taiyuan University of Technology, China

Proceedings Co-chairs

Fu Lee Wang	Caritas Institute of Higher Education, Hong Kong
Ting Jin	Fudan 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 of 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, UK
Renato Iannella	National ICT, Australia
Yan Jia	National University of Defence Technology, China
Yu-Kwong Ricky	Colorado State University, USA
Yoon Joon Lee	KAIST, 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
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 Defence Technology, China
Hui Xiong	Rutgers University, USA
Jun Yan	University of Wollongong, 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 – Part II

Management Information Systems

Mobile Computing

Semantic Web and Ontologies

Web Content Mining

Web Information Classification

Web Information Extraction

Web Intelligence

Web Interfaces and Applications

Web Services and E-Learning

XML and Semi-structured Data

Table of Contents – Part I

Applications of Web Information Systems

Applications of Web Mining

Distributed Systems

e-Government and e-Commerce

Geographic Information Systems

Information Security

Intelligent Networked Systems

Text Clustering Based on LSA-HGSOM*

Jianfeng Wang and Lina Ma

Technology College, North China Electric Power University,
071051 Baoding, China
wjf611@yahoo.com.cn

Abstract. Text clustering has been recognized as an important component in data mining. Self-Organizing Map (SOM) based models have been found to have certain advantages for clustering sizeable text data. However, current existing approaches lack in providing an adaptive hierarchical structure within in a single model. This paper presents a new method of hierarchical text clustering based on combination of latent semantic analysis (LSA) and hierarchical GSOM, which is called LSA-HGSOM method. The text clustering result using traditional methods can not show hierarchical structure. However, the hierarchical structure is very important in text clustering. The LSA-HGSOM method can automatically achieve hierarchical text clustering, and establishes vector space model (VSM) of term weight by using the theory of LSA, then semantic relation is included in the vector space model. Both theory analysis and experimental results confirm that LSA-HGSOM method decreases the number of vector, and enhances the efficiency and precision of text clustering.

Keywords: Text Clustering, Hierarchical GSOM, Latent Semantic Analysis, Vector Space Model.

1 Introduction

With the popularization and application of Internet network has become an important part of the people's working and living, and various search engines have been an indispensable tool to retrieve the necessary resources for the people. However, the Internet search engine can often find thousands of search results. Even if some useful information is obtained, it is often mixed with a lot of "noises" to waste the users' time and money. Therefore, in order to efficiently and economically retrieve the resource subset relevant to the given search request and with the appropriate number, the Text clustering is performed and becomes one of important and hot research fields in data mining[1].

Text clustering is different from Text classification. The latter has them for each category while Text clustering has no category annotates in advance. The Text clustering is to divide the Text sets into several clusters according to the Text contents,

* This research was supported by Young Teachers Research Foundation of North China Electric Power University No. 200811036.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 1–10, 2011.

and requires the similarity of the Text contents in the clusters as great as possible and that of different clusters as small as possible. It can organize the Web Text effectively, but also form a classification template to guide the classification of the Web Text. Therefore, the Text clustering is an important content in the domain of data mining, and also acts as a very important role in text mining. The general procedure of the text clustering methods is as follows. Firstly, the documents to be clustered are transformed into some sets of terms, and term weights are assigned to each term of the sets, then some term weights constitute a feature vector that represents a text. In fact, text clustering means text contents clustering. However, the term sets can not concern with the text contents in clustering process. Therefore, a way of improving text clustering effect is that clustering of documents is based on text conception (or semantic content). The existing methods of text clustering can obtain a single clustering structure; however, the result can not show hierarchical relation among categories. In fact, people need to understand the hierarchical relation among categories. In order to overcome this defect, we present a new method of hierarchical text clustering called LSA-HGSOM method. The method applies the theory of LSA to construct a VSM (Vector Space Model), and achieves text clustering through conception statistic computation. Therefore, the method advances the speed and precision of text clustering. Moreover, the clustering method can achieve automatically hierarchical text clustering through a hierarchical GSOM method (called HGSOM).

2 The Theory of LSA

2.1 Term Matrix

LSA (Latent Semantic Analysis) [2] is one of the most popular linear document indexing methods which produce low dimensional representations using word co-occurrence which could be regarded as semantic relationship between terms. LSA aims to find the best subspace approximation to the original document space in the sense of minimizing the global reconstruction error (the Euclidean distance between the original matrix and its approximation matrix). It is fundamentally based on SVD (Singular Value Decomposition) and projects the document vectors into the subspace so that cosine similarity can accurately represent semantic similarity. Given a term-document matrix $X = [x_1, x_2, \cdots, x_n] \in R^m$ and suppose the rank of X is r, LSA decomposes X using SVD as follows:

$$X = U \sum V^T \tag{1}$$

where $\sum = diag(\sigma_1, \cdots \sigma_r)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the singular values of X. $U = [u_1, \cdots, u_r]$ and u_i is called the left singular vector. $V = [v_1, \cdots, v_r]$ and v_i is called the right singular vector. LSA uses the first k vectors in U as the transformation matrix to embed the original documents into a k-dimensional space.

2.2 Singular Value Decomposition

After the matrix X is established, we can acquire an approximate matrix X_k of the matrix X with k orders, where $k < \min(m,n)$. By the singular value decomposition[3], a matrix X can be denoted as a product of three matrices.

$$X = U \sum V^T \tag{2}$$

In formula (3), U and $\sum$ denote the left and the right singular vector matrices of the matrix X respectively; the diagonal matrix $\sum$ consists of singular values of the matrix X according to the arrangement with descending order. We select the foremost k maximal singular values of the matrix X , and establish an approximate matrix X_k with k order.

$$X_k = U_k \sum_k V_k^T \tag{3}$$

In formula (4), U_k and V_k are orthogonal vectors. X_k denotes approximately the term vector matrix X , the row vectors of U_k represent the term vectors, and the row vectors of V_k represent the document vectors. After using singular value decomposition and selecting approximate matrices of k orders, the model of LSA acquires some good effects as follows. For one thing, the disadvantageous factors in original term matrix are decreased. Moreover, the semantic relation between terms and documents becomes more obvious. In addition, the dimension of VSM is decreased greatly, and so the speed of clustering is advanced. In short, through the process of LSA, the VSM of documents has the following merits.

The dimension of VSM is decreased greatly, and so the speed of clustering is advanced.

3 The Theory of GSOM

3.1 The Self-Organizing Map (SOM)

SOM is an unsupervised neural network model that maps high-dimensional input space to low-dimensional output space. When the resulting map is a two-dimensional topology, the intuitive visualization provides good exploration possibilities. The drawback of this approach is the pre-fixed structure of the output space and lack of providing hierarchical relations between the input spaces [3].

The Growing Self Organizing Map (GSOM).

The GSOM algorithm is composed of three phases, initialization, growing and smoothing. Soon after the smoothing phase, the generated map can be queried and the input data vectors clustered [4].

1) Initialization phase:

Initialize the weight vectors of the starting nodes (usually four) with random numbers between 0 and 1.

Calculate the growth threshold (GT) for the given data set of dimension D according to the spread factor (SF) using the formula $GT = -D \times \ln(SF)$.

2) Growing Phase:

a) Present input to the network.

b) Determine the weight vector that is closest to the input vector mapped to the current feature map (winner), using Euclidean distance. This step can be summarized as: find q' such that $\left|v - w_{q'}\right| \le \left|v - w_q\right| \forall q \in N$ where v, w are the inputs and weight vectors respectively, q is the position vector for nodes and N is the set of natural numbers.

c) The weight vector adaptation is applied only to the neighborhood of the winner and the winner itself. The neighborhood is a set of neurons around the winner, but in the GSOM the starting neighborhood selected for weight adaptation is smaller compared to the SOM (localized weight adaptation). The amount of adaptation (learning rate) is also reduced exponentially over the iterations. Even within the neighborhood, weights that are closer to the winner are adapted more than those further away. The weight adaptation can be described by

$$w_j(k+1) = \begin{cases} w_j(k) & if\ j \notin N_{k+1} \\ w_j(k) + LR(k) \times (x_k - w_j(k)) & if\ j \notin N_{k+1} \end{cases} \tag{4}$$

Where the Learning Rate $LR(k)$, $k \in N$ is a sequence of positive parameters converging to zero as $k \to \infty$. $w_j(k)$, $w_j(k+1)$ are the weight vectors of the node j before and after the adaptation and N_{k+1} is the neighborhood of the winning neuron at the $(k+1)th$ iteration. The decreasing value of $LR(k)$ in the GSOM depends on the number of nodes existing in the map at time k.

d) Increase the error value of the winner (error value is the difference between the input vector and the weight vectors).

e) When $TE_i \ge GT$ (where TE_i is the total error of node i and GT is the growth threshold). Grow nodes if i is a boundary node. Distribute weights to neighbors if i is a non-boundary node.

f) Initialize the new node weight vectors to match the neighboring node weights.

g) Initialize the learning rate (LR) to its starting value.

h) Repeat steps 2 – 7 until all inputs have been presented and node growth is reduced to a minimum level.

3) Smoothing phase

a) Reduce learning rate and fix a small starting neighborhood.

b) Find winner and adapt the weights of the winner and neighbors in the same way as in growing phase.

The growth threshold is based on the number of dimensions of the dataset and the spread factor (SF). SF is a predetermined value in the range 0-1, with zero allowing least spread and one, maximum spread. A limited spread with a smaller SF value should ideally be the starting map. Once significant clusters are identified, they can be used as the basis for further analysis with a higher SF value.

4 Text Clustering Method Based on LSA—GSOM

4.1 Conception of Text Clustering

Text clustering is the process of assigning the documents in a document base to different categories, and is a typical machine learning problem with no supervising. A species is some groups of documents. Documents within one species are more similar than those among different species. Therefore, the aim of text clustering is to group the documents: minimizing the similarity among different species and maximizing the similarity within a species.

Clustering analysis is the process that assigns similar documents to the same categories through computing the degree of similarity among all documents. By the singular value decomposition, the row vectors of v k are the vectors of texts. Therefore, we apply the row vectors of VK to calculating the degree of similarity among documents. The degree of similarity is generally denoted by cosine distance, which is defined as follows:

$$Sim(i,j)=\frac{\sum_{m=1}^{k} W_{im}\times W_{jm}}{\sqrt{\sum_{m=1}^{k}(W_{im})^2\times\sum_{m=1}^{k}(W_{jm})^2}} \tag{5}$$

Where $Sim(i,j)$ is the degree of similarity between text i and j ; where W_{im}, and W_{jm} denote the values of the rows i and j of the column m in the matrix Vk respectively.

4.2 Dimensionality Reduction

Reduction of the data dimensionality may lead to significant savings of computer resources and processing time. However the selection of fewer dimensions may cause a significant loss of the document local neighborhood information. Due to this compromise, we have chosen to use the popular and well studied singular value decomposition.

SVD is used to rewrite an arbitrary rectangular matrix, such as a Markov matrix, as a product of three other matrices: $X=U\sum V^T$. As a Markov matrix is symmetric, both left and right singular vectors (U and V) provide a mapping from the document space to a newly generated abstract vector space. The elements $(\lambda_0,\lambda_1,\cdots,\lambda_{r-1})$ of the diagonal matrix S, the singular values, appear in a magnitude decreasing order. One of the more important theorems of SVD states that a matrix formed from the first n singular triplets $\{U_i,\lambda_i,V_i\}$ of the SVD (left vector, singular value, right vector combination) is the best approximation to the original matrix that uses n degrees of freedom. The technique of approximating a data set with another one having fewer degrees of freedom, known as dimensional reduction, works well, because the leading singular triplets capture the strongest, most meaningful, regularities of the data. The latter triplets represent less important, possibly spurious, patterns. Ignoring them actually improves analysis, though there is the danger that by keeping too few degrees of freedom, or dimensions of the abstract vector space, some of the important patterns will be lost.

After reducing the dimension, documents are represented as n-dimensional vectors in the diffusion space, and can be clustered by using HGSOM.

4.3 HGSOM Clustering

Hierarchical clustering techniques are categorized into agglomerative (bottom-up) and divisive (top-down) approaches[4]. Agglomerative clustering starts with one point clusters and recursively merges two or more similar clusters until all the clusters are encapsulated into one final cluster. Divisive clustering considers the entire dataset as one cluster and then recursively splits the most appropriate cluster until a stopping criterion is achieved. For details on clustering algorithms refer to. The hierarchical clustering model presented in this section builds a hierarchy of clusters in a novel manner. It does not follow a traditional bottom-up or top-down approach, but using GSOM as the basis and utilizing its spread factor. Since the spread factor takes the value between 0 and 1, to avoid missing any significant sub groupings, a set of values across the whole range (0-1) are initialized.

The GSOM uses a threshold value, GT, to decide when to initiate new node growth [5]. GT will decide the amount of spread of the feature map to be generated. Therefore, if only an abstract picture of the data is required, a large GT will result in a map with a fewer number of nodes [4]. Similarly, a smaller GT will result in the map spreading out more. Node growth in the GSOM is initiated when the error value of a node exceeds the GT. The total error value for node i is calculated as:

$$TE_i = \sum_{H_i} \sum_{j=1}^{D} (x_{i,j} - w_j)^2 \tag{6}$$

where H_i is the number of hits to the node i and D is the dimension of the data. $X_{i,j}$ and w_j are the input and weight vectors of the node i, respectively. For a boundary node to grow a new node, it is required that

$$TE_i \geq GT \tag{7}$$

The GT value has to be experimentally decided depending on the requirement for the map growth. As can be seen from (1), the dimension of the data set will make a significant impact on the accumulated error (TE) value, and as such will have to be considered when deciding the GT for a given application. Since $X_{i,j} \geq 0$, $W_j \leq 1$, the maximum contribution to the error value by one attribute (dimension) of an input would be,

$$\max \left| x_{i,j} - w_j \right| = 1 \tag{8}$$

Therefore, from (1)

$$TE_{\max} = D \times H_{\max} \tag{9}$$

where $TE_{\max}$ is the maximum error value and is the maximum possible number of hits. If H_t is considered to be the number of hits at time (iteration) t, the GT will have to be set such that

$$0 \leq GT < D \times H(t) \tag{10}$$

Therefore, GT has to be defined based on the requirement of the map spread. It can be seen from (4) that the GT value will depend on the dimensionality of the data set as well as the number of hits. Thus, it becomes necessary to identify a different GT value for data sets with different dimensionality. This becomes a difficult task, especially in applications such as data mining, since it is necessary to analyze data with different dimensionality as well as the same data under different attribute sets. It also becomes difficult to compare maps of several datasets since the GT cannot be compared over different datasets. Therefore, the user definable parameter is introduced. The SF can be used to control and calculate the GT for GSOM, without the data analyst having to worry about the different dimensions. The growth threshold is defined as

$$GT = D \times f(SF) \tag{11}$$

where $SF \in R, 0 \leq SF \leq 1$, and $f(SF)$ is a function of SF, which is identified as follows.

The total error TE_i of a node i will take the values

$$0 \leq TE_i \leq TE_{\max} \tag{12}$$

where $TE_{\max}$ is the maximum error value that can be accumulated. This can be written as

$$0 \leq \sum_{H} \sum_{j=1}^{D} (x_{i,j} - w_j)^2 \leq \sum_{H_{\max}} \sum_{j=1}^{D} (x_{i,j} - w_j)^2 \tag{13}$$

Since the purpose of the GT is to let the map grow new nodes by providing a threshold for the error value, and the minimum error value is zero, it can be argued that for growth of new nodes,

$$0 \leq GT \leq \sum_{H_{\max}} \sum_{j=1}^{D} (x_{i,j} - w_j)^2 \tag{14}$$

Since the maximum number of hits $(H_{\max})$ can theoretically be infinite, (7) becomes $0 \leq GT \leq \infty$. According to the definition of spread factor, it is necessary to identify a function $f(SF)$ such that $0 \leq D \times f(SF) \leq \infty$

A function that takes the values $0 \rightarrow \infty$, when x takes the values 0 to one, is to be identified. A Napier logarithmic function of the type $y = -a \times \ln(1-x)$ is one such equation that satisfies these requirements. If $\mu = 1 - SF$ and

$$GT = -D \times \ln(1-\eta) \qquad (15)$$

then

$$GT = -D \times \ln(SF) \qquad (16)$$

Therefore, instead of having to provide a GT, which would take different values for different data sets, the data analyst can now provide a value - SF, which will be used by the system to calculate the GT value depending on the dimensions of the data. This will allow HGSOM to be identified with their spread factors and can form a basis for comparison of different maps.

During cluster analysis, it may be necessary (and useful) for the analyst to study the effect of removing some of the attributes (dimensions) on the existing cluster structure. This might be useful in confirming opinions on non-contributing attributes on the clusters. The spread factor facilitates such further analysis since it is independent of the dimensionality of the data. This is very important, as the growth threshold depends on the dimensionality.

5 Experiments

We applied the theory of LSA method to construct the VSM, and applied the HGSOM network to achieve text clustering. We collected 500 documents to carry on the text clustering. These documents were classified into three large-scale species and six child species: Education (Includes campus (CAM) and high school education (HSE)), Economics (Includes industrial (IND) and agriculture (AGR) economics), and Medicine (Includes clinic (CLI) and nurse (NUR)). After the feature selection, we obtained 1863 feature words.

5.1 Experiment 1

We applied directly the words matrix D to carry on the text clustering, and the number of input node was 1863. In this case we could not acquire satisfactory result using HGSOM. The training speed of the network is very slow, and the result of text clustering is not correct. We think that the wrong result comes of too many input nodes. So we could not acquire good result using HGSOM network.

5.2 Experiment 2

In view of the result of the experiment 1, we choose k=785 through the process of LSA. After that, the number of input nodes is reduced to 785. We applied the HGSOM clustering algorithm to the documents clustering. The speed of training is improved greatly, and the results of clustering are shown in the table 1and table 2. We applied average accuracy (AA%)[6] to evaluate the results of text clustering.

In view of the result of the experiment 1, we choose k=785 through the process of LSA. After that, the number of input nodes is reduced to 785. We applied the HGSOM clustering algorithm to the documents clustering. The speed of training is improved greatly, and the results of clustering are shown in the table 1and table 2. We applied average accuracy (AA%)[6] to evaluate the results of text clustering.

In table 1 and table 2, the clustering results are satisfactory; therefore, HGSOM method is feasible for hierarchical text clustering. From table 1 and table 2, we can see that the results of table 1 are better than the results of table 2. The reason of acquiring this result is that the documents have more similar feature words among sub-layers than those documents among large-scale species, which impacts on the result of text clustering.

Table 1. The clustering result of the first layer by hgsom method

Example	(AA)%
Medicine	*88.5*
Education	*90.2*
Economics	*86.3*

Table 2. The clustering result of the first layer by hgsom method

Example	(AA)%
CLI	*79.6*
NUR	*79.1*
CAM	*84.1*
HSE	*82.9*
IND	*85.4*
AGR	*84.9*

6 Conclusions

We can draw the conclusions from above experiments as follows:

（1）The present LSA-HGSOM method is feasible for text clustering.

（2）LSA-HGSOM method can achieve automatically hierarchical text clustering, and overcome the shortcomings of traditional methods.

（3）LSA-HGSOM network is limited for text clustering. If there are many input nodes, we can not acquire good result.

（4）LSA-HGSOM method can enhances the efficiency and precision of text clustering.

In this paper, we proposed a new text clustering method LSA-HGSOM. In this method, it firstly makes preprocess of texts, and introduced LSA theory to improve the precision of clustering and reduce the dimension of feature vector. Then it used HGSOM to execute the clustering to the texts, In the experiment, the result shows that LSA-HGSOM method would get a better effect in the text clustering.

References

1. Lawrence, R.D., Almasi, G.S., Rushmeier, H.E.: A scalable parallel algorithm for selforganizing maps with applications to sparsedata mining problems. Data Mining and Knowledge Discovery 3, 171–195 (1999)
2. Dumais, S.T.: Using latent semantic analysis to improve information retrieval. In: CHI 1988 Proceedings, pp. 281–2853 (1988); Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
3. Rauber, A., Merkl, D.: Using self-organizing maps to organize document archives and to characterize subject matters: How to make a map tell the news of the world. In: Bench-Capon, T.J.M., Soda, G., Tjoa, A.M. (eds.) DEXA 1999. LNCS, vol. 1677, pp. 302–311. Springer, Heidelberg (1999)
4. Hsu, A., Tang, S.-L., Halgamuge, S.K.: An unsupervised hierarchical dynamic self-organising approach to cancer class discovery and marker gene identification in microarray data. Bioinformatics 19, 2131–2140 (2003)
5. Amarasiri, R., Alahakoon, D.: Applying Dynamic Self Organizing Maps for Identifying Changes in Data Sequences. IEEE Transactions on Neural Networks, Special Issue on Knowledge Discovery and Data Mining 11(3), 601–614 (2000)
6. Jiang, N., Shi, Z.-z.: Bayesian posteriori model selection for text clustering. Journal of Computer Research and Development, 5 (2002)

Design Pattern Modeling and Implementation Based on MDA

Xuejiao Pang[1], Kun Ma[2], and Bo Yang[1]

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing,
University of Jinan, Jinan, China
{nic_pangxj,yangbo}@ujn.edu.cn
[2] School of Computer Science and Technology, Shandong University, Jinan, China
nic_makun@ujn.edu.cn

Abstract. Model Driven Architecture (MDA) stresses on the model-centric. It defines the framework of the system by using various models. Aiming to increase not only the modeling granularity but also the reusability of model transformation rule we apply the design pattern into MDA. In this paper, firstly, a modeling approach based on role is presented. In this way, the pattern model and the transformation rule can be defined respectively. Secondly, two extended meta-meta-models, ExPattern(Extended Pattern) and ExRole(Extended Role), which are the meta-models of Pattern and Role respectively, are demonstrated in the article. A QVT-based transformation rule is defined for the snake of models transformation. At last, a case study of Graduate Education Management System which uses the technologies proposed in this paper is demonstrated.

Keywords: design pattern; model transformation; MOF; meta-model; QVT.

1 Introduction

Developers today have to cope with various challenges dealing with design of collaborative system because of some complex characters of collaborative system [1]. Model Driven Architecture (MDA), which is a new approach to software development put forward by OMG (Object Management Group), realizes the separation between the design of the system and its implementation. It ensures the system can be executed on various platforms without changing the design. The shift of the development focus from the code to the model is a significant aim of MDA. Unlike traditional software development which tends to be code-centric, MDA stresses on life-cycle of the software development [2]. Modeling plays a key role in MDA. There are two kinds of MDA modeling: PIM (Platform Independent Model), PSM (Platform Specific Model). PIM is an abstract specification of the software business logic, the behavior of the application cannot be modeled in detail. In this case it is easier to write source code manually [3]. PSM produced by the transformation is a model of the same system specified by the PIM, it also specifies how that system makes use of the chosen platform [4]. Therefore, the model transformation can be considered as a mapping between the models via using transformation rule [13].

The rest of the paper is organized as follows: In Section 2 the related work is introduced. In this section, first we describe the way how to introduce the design

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 11–18, 2011.

patterns into MDA briefly, second a modeling method which is based on role is provided. Section 3 presents the various meta-models of the design pattern proposed in the paper. In this section, firstly, the RoleOf relationship is presented. Secondly, the extended MOF meta-meta-model, which defines the pattern specifications is demonstrated. Section 4 discusses the model transformation approach based on QVT. In Section 5, a case study of Graduate Education Management System uses MDA and its primary technologies is described. At last, conclusions are shown in Section 6.

2 Related Work

The purpose of this paper is to increase not only the modeling granularity but also the reusability of model transformation rules by using MDA based on design patterns [14]. It is different from the former work related to the design pattern.

2.1 The Design Pattern

In recent years, object-oriented design pattern has been widely used in the development of software systems, as systems become increasingly complex and hard to maintain [5]. Design pattern, which provides reusable constructs, can be introduced into MDA to increase the modeling granularity as well as the integrated development units in MDA. This ensures systems to be extensible and flexible since we cannot know all the requirements and build a perfect system at the beginning [6].

2.2 MOF

MDA requires that in order to enable the transformation and exchange, the modeling languages which is used to PIM and PSM should be conformed to MOF.

MOF is defined as a four-layered architecture described as follows:

From the top to the bottom, the four-layered architecture can be considered as M3, M2, M1 and M0, and each level is the instance of the upper level as well as the abstraction of the lower level.

The M0 layer, on which the instance of system can run, consists of the concrete object and the data, is the information layer; The M1 layer, comprised by various models, is the model layer. The model on M1 is used to describe the data of M0. The instance of UML meta-models is demonstrated on M1 layer; The M2 layer, which consists of meta-meta data, is the meta-model layer. The meta-model is the abstraction of the model. UML meta-model is defined on M2 layer. The M3 layer, which defines the meta-model on M2 layer, is the language used by MOF to build lower level meat-models. This layer is the meta-meta-model layer.

2.3 QVT

In response to the need for a standard approach to define mapping functions that map between models, OMG issued the MOF 2.0 Query/Views/Transformation (QVT) Request for Proposals (RFP) [7].

The QVT specification provides a hybrid transformation language, including: Relations, Core and Operational Mappings. The three transformation languages provide the declarative and imperative transformation constructs. The languages Relation and Core can be used to specify the declarative transformation, QVT provides two options to extend declarative specifications with imperative transformation constructs, the Operational Mappings and Black Box operation [8].

3 The Meta-model of the Design Pattern

3.1 The Meta-Model of RoleOf Relationship

Generally, the application class is bound with the role class during the pattern instantiation. But in this way, barriers are still existing, such as pattern overlapping, troubles in reusing the pattern code and so on. In order to solve the problems above, the RoleOf relationship proposed in [11] was applied. In this way, roles are treated as the independent modeling elements and the RoleOf relationship is used to associate a role with an application class as shown in figure 1.

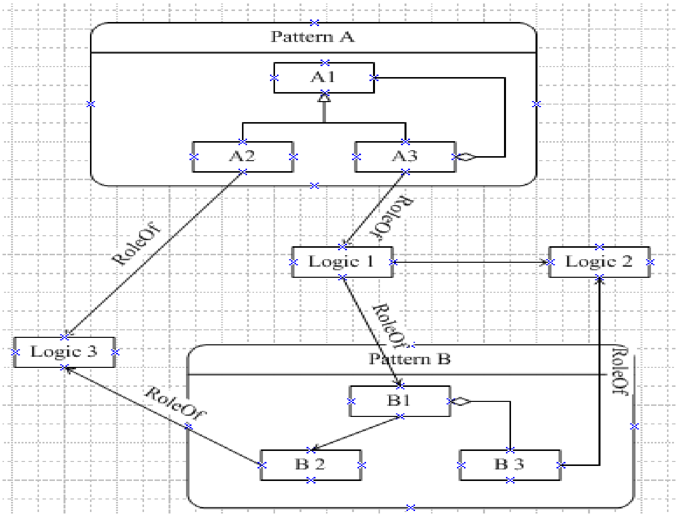


Fig. 1. Binding of Business Class and Rule. The role Ai is associated with Logic j by using the RoleOf relationship. It realizes the separation between the business logic and the model logic.

But this method does not completely realize the separation between the application class and role class. So the ExRole is presented to solve the barriers above.

3.2 The Meta-model of the ExPattern Unit

Each pattern model is the instantiation of the pattern meta-model. ExPattern (Extended Pattern) defines the edge of the pattern meta-model. As figure 2 shows, pattern unit is constituted by the pattern-name and roles.

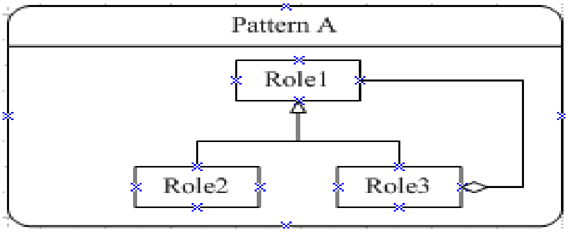


Fig. 2. The Meta-model of Pattern Unit. Pattern unit is constituted by the pattern-name and roles.

3.3 The Meta-model of ExRole

ExRole (Extended Role) defines the meta-models of the participant. Each participant is related to the type, the type can be class or interface. The role describes the structure of the pattern and the responsibility of the participant. Class, attribute of class, operation of class and so on are all treated as role. This effectively solves the difficulties during instantiating the design patterns.

As figure 3 shows, each participant is inherited from the role class. The participant can be the class role, attribute role, operation role, parameter role and relation role. That solves the problems during instantiation of the pattern.

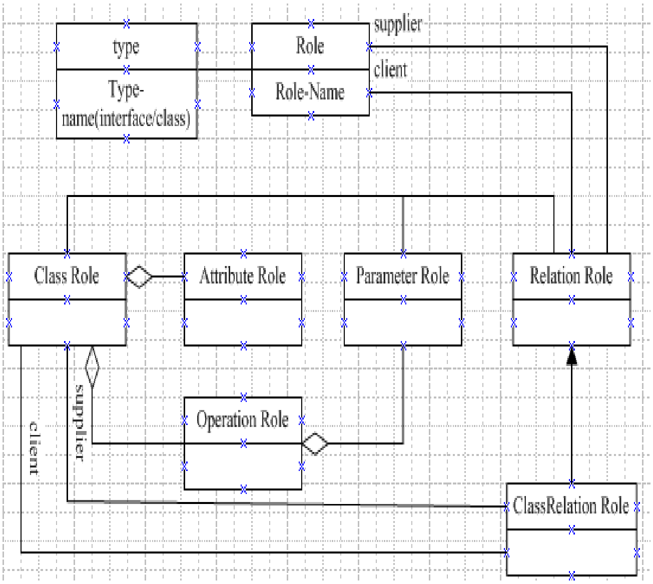


Fig. 3. The Meta-Model Of the Role

3.4 The Meta-model of the Observer Pattern Unit

Reference [11] introduces a figure element system based on Observer pattern. The observer pattern is a common way of design pattern, it realizes the consistency between the associated objects. It is defined by GoF as which an object maintains a list of its observers and automatically notifies them of any state changes, usually by calling one of their methods [10]. The observer pattern includes two roles, the subject and the observer (figure 5). As shown in figure 4, each subject maintains a list of observers, it implements an interface to notify the observers when changes occur in the subject. Each observer can observe any subjects it is interested in at any given time and receive the update notification from the subject.

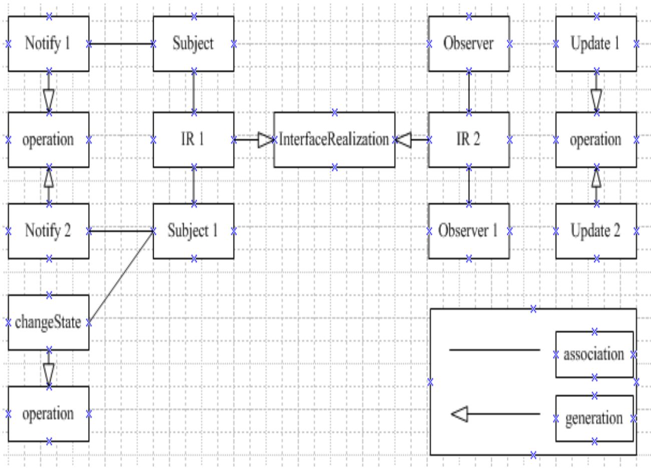


Fig. 4. Pattern Specification of Observer Pattern. It describes the pattern specification of observer pattern on M2 layer.

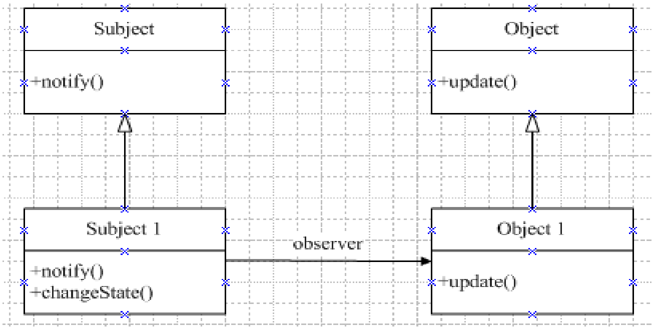


Fig. 5. An Instance Of Observer Pattern. It describes an instance of observer pattern.

4 The Transformation Rule

In this paper, we apply the QVT-based transformation rule.

The input to the transformation is the marked PIM and the mapping, the result is the PSM and the record of the transformation [4]. For the specification of evaluation rules we use the formalism of model transformations, more precisely a graphical syntax of

QVT rules [9]. Firstly, the transformation tool acquires the source model, secondly it implements the transformation definition, and gives the target model as output. When the QVT rule is implemented on the source model, a LHS matching sub model of the source model is searched and a target model is obtained by rewriting the matching sub model by a new sub model that is derived from RHS under the same matching [9].

5 Case Study

The model driven development is demonstrated in detail via a case study of the Graduate Education Management System. In this section, the modeling of PIM is introduced and a QVT-based transformation definition is used to mapping PIM to PSM.

The Graduate Education Management System displays kinds of information to different participant, including: student, teacher, college secretary and super administrator. Take the process of course selection for example. When a student selects a new course, the list of selected course, the student list which describes who has selected the course, the achievement list which records the score of students selected this course and other data which is associated with the information of select course are all needed to be update. When the data is changed, data layer sends the update request, and the interface layer receives the update message and shows the updated information according to the different request of the view.

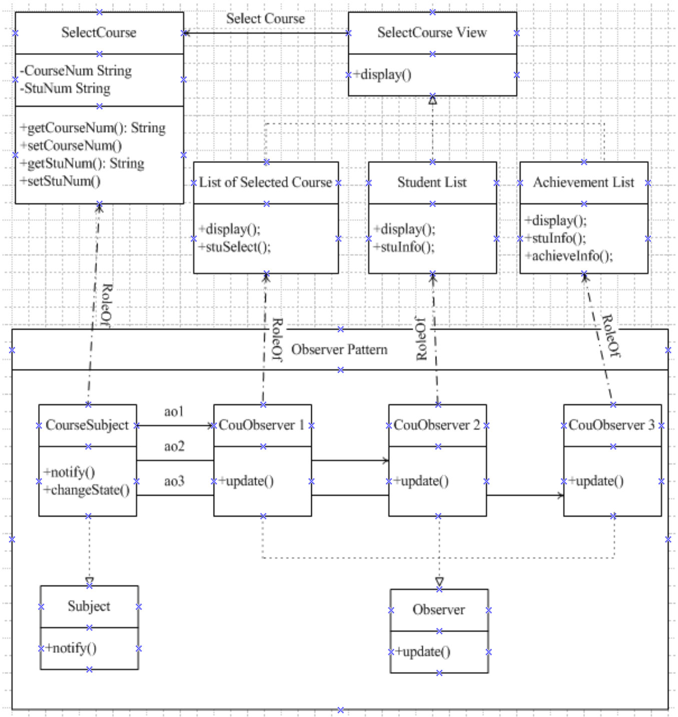


Fig. 6. Binding of Business Model and Observer Pattern

5.1 Modeling of PIM

According to the request, the view should reflect the latest information, so the observer pattern is applied. In this paper, we use the RoleOf relationship to realize the business models and pattern models separated, firstly we construct the pattern model and business model, secondly the pattern models and the business models should be associated by applying the RoleOf relationship. Figure 6 presents the binding of business model and observer pattern model.

As is apparently demonstrated in figure 6, a CourseSubject covers three classes of Observer, including: CourseObserver1, CourseObserver2, CourseObserver3. And the SelectCourse, StuOfSelectCourse and StuAchievement belong to the business meta-model. Then we should bind the instances of CourseObserver1, CourseObserver2 and CourseObserver3 with SelectCourse, StuOfSelectCourse and StuAchievement respectively.

5.2 The Transformation of PIM-PSM

When the modeling of PIM is finished, PSM should be generated as follows:

Firstly, the model should be checked whether is associated with RoleOf relationship. If it is associated with RoleOf then the transformation rule proposed in this paper can be executed. Secondly a temporary intermediate variable used to save the combined models should be defined, this variable would be released before the ending of the transformation. When the rule is implementing, the operations and attributes in the business models would be appended to the role. At last, according to the type of the role, interface or class, the object in the source model can be transformed to the corresponding type of target model.

5.3 Analysis of the Effect

The method of applying design pattern in MDA increases the modeling granularity as well as the reusability of model transformation rules. The RoleOf relationship solves the barriers such as pattern overlapping and traceability during the pattern instantiation, which improves the reusability of the models.

By instantiating the observer pattern we realize the modeling based on the observer pattern units. By separating the business model and logic model clearly, it can solve the traceability and reusability of the model to a large extent.

6 Conclusion

In this paper design pattern is introduced into MDA. A modeling approach based on role is proposed. In addition, this paper shows how to extend MOF meta-meta-models to define the specification and how to define the model transformation rules based on QVT. Finally, the model driven development process of Graduate Education Management System is demonstrated to verify the approach.

Next, we will consider increase the complementary transformation rules to support the comprehensive automatic transformation of system.

Acknowledgment. This work was supported by the Provincial Natural Science Foundation for Outstanding Young Scholars of Shandong under Contract Numbers JQ200820 and the Program for New Century Excellent Talents in University under Contract Numbers NCET-10-0863. We would like to thank the anonymous reviewers for valuable comments.

References

1. Xiang, Y., Zhang, S., Shi, M.: Boosting creativity of CSCW research:survey and trend analysis. Journal on Communications 27(11), 1–6 (2006)
2. Hamous-Lhadj, A., Gherbi, A., Nandigam, J.: The Impact of the Model-Driven Approach to Software Engineering on Software Engineering Education. In: 2009 Sixth International Conference on Information Technology: New Generations, pp. 719–724 (2009)
3. Chen, Z., Ma, K., Abraham, A., Yang, B., Sun, R.: An Executable Business Model for Generic Web Applications. In: Proceedings of International Conference on Computer Information Systems and Industrial Management Applications, Kraków, Poland, pp. 573–577 (2010)
4. Miller, J., Mukerji, J.: MDA Guide (2003)
5. Zhao, C., Kong, J., Zhang, K.: Design Pattern Evolution and Verification Using Graph Transformation. In: Proceedings of the 40th Hawaii International Conference on System Science, pp. 1530–1605 (2007)
6. Cinneide, M.O., Nixon, P.: Automated Software Evolution Towards Design Patterns. In: Proceedings of the 4th International Workshop on Principles of Software Evolution, pp. 162–165 (2001)
7. Object Management Group. Request for Proposal: MOF 2.0 Query/View/Transformation RFP. OMG (2002), `http://www.omg.org/docs/ad/02-04-10.pdf`
8. Romeikat, R., Roser, S., Mullender, P., Bauer, B.: Translation of QVT Relations into QVT Operational Mappings. Computer Science, 137–151 (2008)
9. Marković, S., Baar, T.: Semantics of OCL specified with QVT. Software and Systems Modeling, 399–422 (2008)
10. Liu, J., Yin, H., Wang, Y.: A Novel Implementation of Observer Pattern by Aspect Based on Java Annotation. Computer Science and Information Technology, 284–288 (2010)
11. He, C., He, K.: A Role-Based Approach to Design Pattern Modeling and Implementation. Journal of Software 17(4), 658–669 (2006)
12. Ma, K., Yang, B., Chen, Z., Li, Q., Cui, L.: Research of Model-driven Web A|pplication Rapid Development Platform. Computer Science 37(11), 29–33 (2010)
13. Ma, K., Yang, B.: A Hybrid Model Transformation Approach Based on J2EE Platform. In: Proceddings of 2nd Internation Workshop on Education Technology and Computer Science, ETCS 2010, China, Wuhan, pp. 161–164 (2010)
14. Zhang, T., Zhang, Y., Yu, X., Wang, L., Li, X.: MDA Based Design Patterns Modeling and Model Transformation. Journal of Software 19(9), 2203–2217 (2008)

End-to-End Resources Planning Based on Internet of Service

Baoan Li and Wei Zhang

Computer School, Beijing Information Science and Technology University
Beijing, P.R. China
liba2010@139.com

Abstract. Business innovation and collaboration lead to the development of enterprise information systems from the internal scope towards the more broadly external parties. EERP (End-to-End Resources Planning) as the up to date requirement in the enterprise information field is presented in this paper. The relevant technologies about internet of service to realize EERP have been researched in detail. An approach based on value-aware service engineering and service network technologies for EERP has been proposed and applied, which can help to realize business-goal-driven dynamic semantic integration of the Web services.

Keywords: Internet of Service, SOA (Service Oriented Architecture), EERP (End-to-End Resources Planning), Service Network.

1 Introduction

At present, enterprises are in the period of social transformation of backward industry, their main characteristics into service as the dominant social activities, and information technology as the supporting technology of community, business and personal activities. At the same time, with the rapid development of e-Business, Web applications based on the Web are developed from localization to globalization, from B2C (Business-to-Customer) to B2B (Business-to-Business), from centralized fashion to decentralized fashion. Web service is a new application model for decentralized computing, and it is also an effective mechanism for the data and service integration on the Web [1]. It is important and necessary to carry out the research on the new architecture of web services, on the combinations with other good techniques, and on the integration of services. It has started from information integration, process integration to enterprise integration, business integration. It has become an inevitable trend to combine the existing business units, service, and application into one in order to meet user demand.

The traditional ERP (Enterprise Resources Planning) systems which focused on optimization of the enterprise internal process are not adapted to the requirement for interaction possibilities with external parties. EERP (End-to-End Resources Planning) is initiated to serve this purpose.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 19–26, 2011.

The relevant technologies and methodologies about Internet of Service such as Service Computing, Service Engineering and Service Network have been proposed. In especial, SOA and component oriented technologies has been succeed in many fields. The approach of component oriented technology even became the recommended approach of the road map of SOA in China [2].

It is very urgent and important to use the new technologies about Internet of Service into the development of End-to-End Resource Planning.

2 The Development of Enterprise Information System

Enterprise information system has been developed from MRP (Material Requirements Resource Planning), MRPII (Manufacturing Resource Planning), ERP (Enterprise Resource Planning) and ERPII (Enterprise Resources Planning and Co-Business) towards EERP (End-to-End Resources Planning).

The development of enterprise information system shows as Fig. 1. Emphatically, the relations amongst them are the kind of development or including, but not replacement or negation.

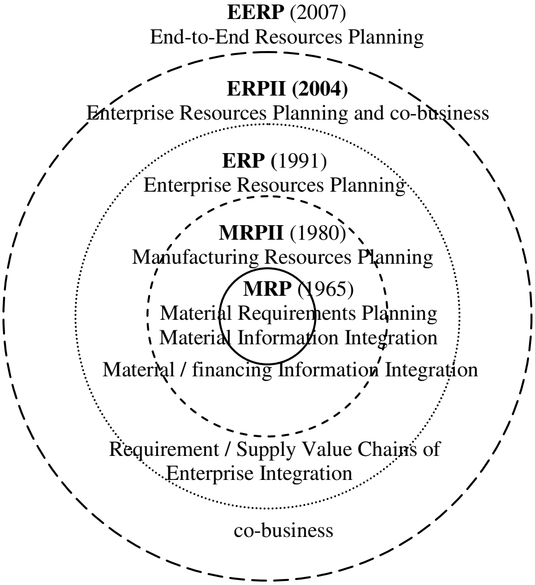

EERP (2007)
End-to-End Resources Planning
ERPII (2004)
Enterprise Resources Planning and co-business
ERP (1991)
Enterprise Resources Planning
MRPII (1980)
Manufacturing Resources Planning
MRP (1965)
Material Requirements Planning
Material Information Integration
Material / financing Information Integration
Requirement / Supply Value Chains of
Enterprise Integration
co-business

Fig. 1. The Development of Enterprise Information Systems

The development of ERP and EERP can be also described as changing from the centre of resources towards the centre of relation [3]. SOA, SSOA (Semantic SOA) and Component oriented technology offer the new way to accomplish ERP/EERP on the relation platform.

3 The Relevant Technology and Methodologies on Internet of Service

3.1 Service Computing and Service Science

Service computing or service science as a new research field has gained more and more attention. It has passed through two development stages [4]:

At first stage, Garter Group proposed the concept of Service Oriented Architecture (SOA) in 1996, to make service computing development rapidly. Service computing appeared the first high tide. Service oriented programming paradigm' decoupling, based on open standards interoperability, large particle reuse, supporting dynamic expanding technologies have begun enjoys popular support. More and more projects have begun to use SOA methodology in EAI (Enterprise Application Integration) and other application fields such as End-to-End resource planning to seek the software reuse, flexibility, low cost and rapid development.

The development processes of SOA are also divided into three phases [5]:

At first SOA focuses on the integration in enterprise and resolves the One to One relations. It has started from 2003 year.

Second SOA phase focuses on the value chains between the credible associate enterprises and resolves the One to Many relations. It has begun from 2007 year.

The third phase of SOA focus on finding new associate and new services. It resolves the Many to Many relations. However, the start year can be unconfirmed. It depends on the advanced research and application about SOA and other new technologies.

At second stage of service computing, IOT (the Internet of Things), Social Information Network and Cloud Computing have gradually become the most concern focus. SOA, SaaS (Software as a Service) and SOC (Service Oriented Computing) represent the general trend of the future. The development of service computing is entering into the second high tide. It mainly reflected in two aspects:

One is software and resources are put in the cloud and as the infrastructure, and then the consumers need not setup or deploy them on their local computers in many cases. Another is the software using and operating mode about XaaS (Anything as a Service) will support users to use rather than owning, to consume and use information and communication technology resources with pay-on-demand mode. Service is not only the link or adhesive among the infrastructure and the user experiences but also the kernel carrier of the various kinds of the exposed intelligence in new network environments with dynamic, open, indeterminacy and assembly characters.

3.2 Values-Aware Service Engineering Methodology

VASEM (Value-Aware Service Engineering Methodology) uses the model driven idea to transform from top to bottom, and payes equal attention to service value and service function. In the transformation process, if existing more aspects to influence service values then make decision by the value-aware method [6].

The implementation processes of VASEM shows as Fig.2. Left side in Fig.2 is service model space. It started from the design of service scheme, to service modeling, and then to service system implementation. And right side is service value space. It includes multilevel value models (Value Declaration Model, Value Network Model and Value Dependence Model). Establishing the connections between values and service functions by value tagging, and then to form Value Tagging Models. Therefore analysis and optimizing about service model can be done by the value oriented method. When all values can be fully supported by service models then enter the service system implementation process.

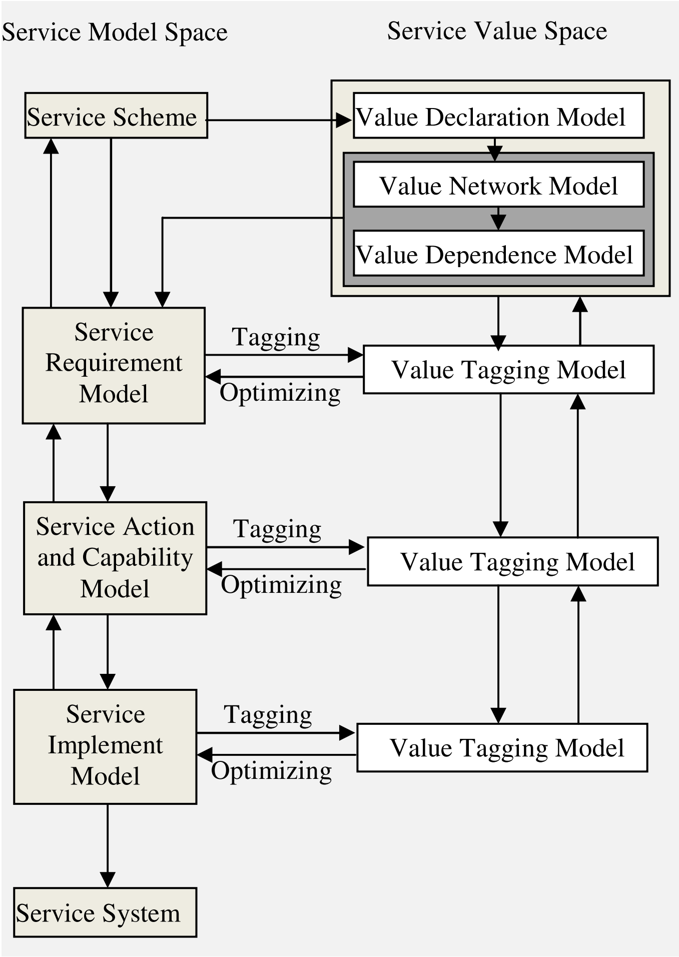


Fig. 2. The Implementation Processes of Value-Aware Service Engineering Methodology

There are some kernel links in VASEM: Value Modeling, Value Tagging, Value Oriented Analysis, Value Oriented Model Optimizing, Feedback and Redesign, Value Oriented Combination and System Implementation.

3.3 Service Network

SN (Service Network) as an infrastructure of Internet of Service offers supporting to the SOC (Service Oriented Computing). It can be as the Social Network with available services on internet. Its kernel is to resolve the services which have rich semantic information and the service interactive relations. Fig. 3 gave the architecture of Service Network [7].

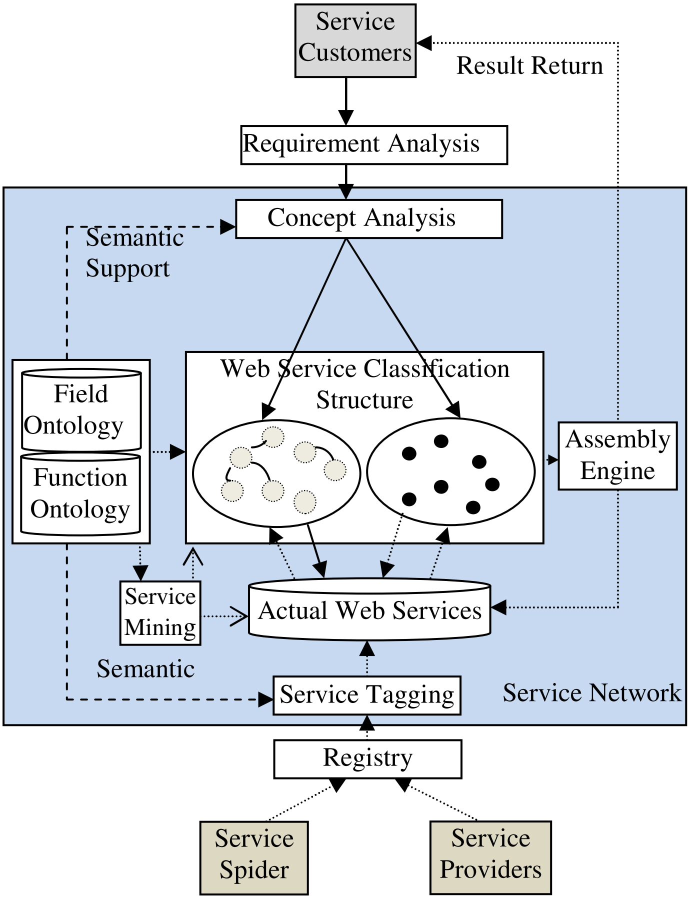


Fig. 3. The Architecture of Service Network

SN can be described as:

SN (V, E) = Graphws<Vws, Ews>, there into

Vws = {Abstract Services} ∪ {Actual Services}

Ews = SRType < Vws, V’ws >

SN is a Social Graph, its nodes are services and edges are service relations. The service nodes are the ontology models with well-defined characters, and can be divided into abstract services and concrete services. The service nodes consist of SN by the semantic relations.

4 Research on EERP

EERP is a business-centric approach to end-to-end integration and optimization of business processes and services to increase business agility and adaptability to ever-changing environment, improve business performance, sustain competitive advantage, and ultimately create business value and accomplish the business objective.

Here, the end-to-end business processes mean the business processes span lines of business, tapping into functional tasks within each of them and beyond the boundaries of the enterprise to trading partners and external service providers. They are the primary means of defining the business and creating differentiating value. To keep these End-to-End business processes in line with the ever-changing demands of the market, we need a deliberate, thorough approach to managing them.

EERP system consists of services that can be evaluated, optimized and replaced. It adapts to the ever-changing requirements, to speed up the establishing progress and reused programming.

EERP is a base platform which is supported by the knowledge bases such as FIPA ontology [8], SCM ontology [9] and M3PO ontology [10] based on SSOA (Semantic SOA).

Firstly, the most adapted business process solution can be found in EERP through the simulating business management engine. Secondly, it can process the exceptions and the faced questions automatically with dynamic responding. The main core technology of EERP includes: SOA, BPM (Business Process Management) and semantics.

There are several main attention service requirements of EERP as following:

4.1 Service Issuance

This function issues the services in the service register library, and then to offer the detailed introduction about service function and QoS (Quality of Service).

4.2 Service Finding

This function is used to locate the appropriate services according to the business goal.

4.3 Service Selection

This function is used to select the appropriate services according to the business goal.

4.4 Service Consulting

Its function is to consult with the service provider and the service consumer, and then to make terms at QoS and SLA (Service-Level Agreement), including the exception processing and the compensating mechanism.

4.5 Service Assembling

This function assembles and binds the services so that to implement the business goal of service consumers.

4.6 Service Harmonizing

It is used to resolve the harmonizing problem among the protocol, data, process and application, and then to procure the optimization result.

4.7 Process Simulation

Before deploying the business process, it arranges the correlative properties and runs at the BPM layer analogously, so that to find the potential problems. The analogous data can be gotten from the history records. This will help the business analyst to reconstruct the processes.

4.8 Process Implementing

It can connect and employ the services, and process the exceptions dynamically through the service finding, service selection and service consulting.

4.9 Process Optimization

It can find the most effective business process through analyzing, monitoring and simulating procedures on BPM (Business Process Management) layer.

5 Conclusion

End-to-End Resource Planning is becoming the new application field of enterprise information systems. Its development must be connected to the new technologies and methodologies on Internet of Service. In order to ensure the project implement successful, it is very important to define and issue the services at the basis of analyzing the value points and value-chains of actual management actives [11]. Assembling services and implementing business process need to use the value-aware service methodology. With the research and application on Internet of Service towards more and more deeply, the relevant standards such as SCA (Service Component Architecture) [12] and SDO (Service Data Object) [13] standards, and SSOA (Semantic SOA), Service Network and other relevant technologies become more mature, End-to-End Resource Planning will be applied widely and greatly increase the competition ability of enterprises in future.

Acknowledgments. The 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 also supported by Funding Project for Natural Science Foundation of China (Grant No. 61070119) and Scientific Research Funding Project for Beijing Information Science and Technology University (Grant No. 1125017). Especially thank Professor Li Ning and Professor Zhang Yangsen who work in the Computer School of Beijing Information Science and Technology University, and the experts in Beijing ChangFeng Open Standards Platform Software Alliance. We also thank the experts and the organizations that appeared in the paper references.

References

1. Yue, K., Wang, X., Zhou, A.: Underlying Techniques for Web Services: A Survey. Journal of Software 15(3), 428–429 (2004)
2. IDC: China road map of SOA. IDC White Paper (2007), http://gocom.primeton.com/special/soabook/soa.php.
3. Li, B.: Research on the production scheduling management system based on SOA. In: Wang, F.L., Gong, Z., Luo, X., Lei, J. (eds.) Web Information Systems and Mining. LNCS, vol. 6318, pp. 286–294. Springer, Heidelberg (2010)
4. Han, Y., Xu, X., He, K.: Service Computing for the Future Internet. Journal of Communication of China Computer Federation 6(9), 10–11 (2010)
5. Li, B.: An Approach to Build Information System Based on SOA and Component Oriented. In: Proc. the IEEE International Conference on Networking and Digital Society, pp. 661–664 (2010)
6. Xu, X., Wang, Z.: Value-Aware Service Model Driven Architecture and Methodology. In: Proc. the 20th IFIP World Computer Congress, pp. 277–286 (2008)
7. Feng, S., Chen, S., Wang, H.: Service Network. Journal of Communication of China Computer Federation 6(9), 26–28 (2010)
8. FIPA nonmadic application support ontology specification (2001), http://www.fipa.org/specs/fipa00014/SI00014H.pdf
9. Haller, A., Gontarczyk, J., Kotinurmi, P.: Towards a complete SCM ontology – The case of ontologizing RosettaNet (2007)
10. Izza, S., Vincent, L., Burlat, P., et al.: A unified Framework for Enterprise Integration – An Ontology – Driven Service Oriented Approach (2007)
11. Li, B., Zhou, W., He, Y.: To Design Component Oriented ERP System Based on Analyzing Enterprise Value-chains. Journal of Computer Engineering and Design 29(15), 3927–3928 (2008)
12. Barack, R., et al.: SCA Java Component Implementation Specification (2007), http://www.osoa.org/display/Main/Service+Component+Architecture+Specifications
13. Adams, M., et al.: Service Data Objects For Java Specification, http://www.osoa.org/display/Main/Service+Data+Objects+Specifications (2006)

A Comprehensive Reputation Computation Model Based on Fuzzy Regression Method of Cross-Domain Users

Juan Zhou[1], Gang Hu[1], and Qinghua Pang[2]

[1] Information Center, Hohai University
[2] Business School, Hohai University
213022 Changzhou, China
zfh8824_cn@sina.com

Abstract. It is a problem that users' reputation distributed in each domain can't be shared in collaborative environment of cross-domain. A computation model of comprehensive reputation based on fuzzy regression is proposed in this paper to supply uniform reputation evaluation method which is the base of collaboration of cross-domain users. The model builds up user's reputation vector of single domain, and the vectors are combined according to weighting coefficients. In order to build up fuzzy regression model, the comprehensive reputation is hazed with symmetric triangular fuzzy number to determine fuzzy coefficients, and to calculate out the result. The experiment results show that the model reflects the flexible range objectively.

Keywords: Reputation Computation, Cross-Domain, Comprehensive Reputation, Fuzzy Regression.

1 Introduction

With the continuous deepening of network application, it's a more and more pressing need of interaction between domains. Users usually access some relatively regular AS domains besides the domains themselves belong to. For example, in CERNET[1], colleges forms AS domains with their teachers and students as relatively steady user groups. Teachers and students need to visit some applications in certain other colleges' domains for cooperation and resource sharing. And another case lies in huge distributed, anisomery and open grid environment, whose users usually visit inter-organizational shared resources.

Certainly the cross-domain cooperation greatly improves Internet resource utilization rate to increase quality of service. However how to manage the access from cross-domain users efficiently is a big challenge. This paper researches a model based on fuzzy regression to calculate users' comprehensive reputation to provide uniformed reputation evaluation method for the convenience of identity management, restrain user's activity, decrease destructive behavior and network disorder. It also supplies a judgment basis for other users' behavior decision. In classical regression analysis, the deviation between evaluated value and actual value is thought as being caused by random error, however in fact, it may be caused by artificially defined structure or inaccurate observation. So the fuzzy regression is applied in the research.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 27–34, 2011.

2 Calculation Model

2.1 Analysis of Reputation in Single Domain

There are variable reputation computation methods in Internet since variable applications have their own characteristics. In this model, multi-dimensional vector is adopted to describe user's reputation to avoid the awkward of distinct reputation content of different application. Each dimension denotes specific application category according to actual need such as e-commerce, data translation and BBS. The domain a user lies in is called primary domain of the user, while the other domains he visits are defined as active domains. One category of application corresponds to only one dimension, and there may be several categories of application in a single domain.

Definition 1. $(T_1^i, T_2^i, ..., T_S^i)$ is a user's reputation corresponding to S categories of application systems in domain i. In the same domain i, reputation records are sampled at q different time stages of each dimension of the same user to form user's reputation matrix in single domain: $\begin{pmatrix} T_{11}^i & T_{12}^i & T_{13}^i \cdots & T_{1q}^i \\ T_{21}^i & T_{22}^i & T_{22}^i \cdots & T_{2q}^i \\ \vdots & & & \\ T_{S1}^i & T_{S2}^i & T_{S2}^i \cdots & T_{Sq}^i \end{pmatrix}$. The updating frequency of user's reputation of different application can be quite different according to different standard. So each application should be considered for the selection of sampling time stage to reflect the updating frequency reasonably. If there is no record at time stage j(q>=j>1), the record is replaced with record at j-1, and the like, if there is no record at stage 1, the user should have never used this type of application. It's suitable to set a neutral value but not 0 as the initial value that not only provides chance for users with good reputation to be served, but also restrains evil behavior to increase the cost of cheating and strike evil users' enthusiasm. It's thought as impossible to appear that most users have no reputation record [2].

2.2 Comprehensive Reputation Computation Model of Cross-Domain

Reputation Conversion of Cross-Domain

In order to calculate the comprehensive reputation of cross-domain, it's necessary to combine the scattered reputations both in primary and active domains. Equation (1) is user reputation from the weighted reputation vectors of each single domain.

$$R_{jq} = \sum_{i=1}^{D_N} \left[W^i * \left(T_{jq}^i \right) \right] \tag{1}$$

R_{jq} represents statistic reputation value of dimension j at time q, j=1…S. W^i represents weighting factor of the user reputation in domain i. T_{jq}^i represents the sampling value of dimension j at time q in domain i. There is relation between user's reputation and activity degree, because if a user visits certain domain more frequently, he concerns it more and his behavior is closer to his real reputation than in other domains. So here W^i is decided by frequency of visitation in each domain, i.e. W^i comes from the value of visit times in domain i divided by the sum of visit times of the user in all the domains. Comprehensive reputation of all past time stages can be expressed by the following equation set:

$$\begin{cases} p_1 = A_0 + A_1 R_{11} + A_2 R_{21} + \cdots + A_S R_{S1} + \delta_1 \\ p_2 = A_0 + A_1 R_{12} + A_2 R_{22} + \cdots + A_S R_{S2} + \delta_2 \\ \vdots \\ p_q = A_0 + A_1 R_{1q} + A_2 R_{2q} + \cdots + A_S R_{Sq} + \delta_q \end{cases} \tag{2}$$

p_j is the total value calculated from reputations of each dimension at time j($1 \le j \le q$), $A_0, A_1, A_2, \cdots, A_S$ are regression coefficients to be determined, $\delta_1, \delta_2, \delta_3, \delta_4$ are random error variables obeying normal distribution of N(0, σ^2). Set $p = (p_1, p_2, \cdots, p_q)^T$, $A = (A_0, A_1, A_2, \cdots, A_S)^T$, $R = \begin{pmatrix} 1 & R_{11} & R_{21} \cdots & R_{S1} \\ 1 & R_{12} & R_{22} \cdots & R_{S2} \\ \vdots & & & \\ 1 & R_{1q} & R_{2q} \cdots & R_{Sq} \end{pmatrix}$, $\delta = (\delta_1 \quad \cdots \quad \delta_q)^T$, the relation between comprehensive reputation and each dimensional reputation represented in the equation set (2)can be written as: $p = R * A + \delta$.

Model Based on Fuzzy Regression

The paper predicts comprehensive reputation with fuzzy mathematics for the uncertainty of user's reputation. The weighted sum of multi-dimensional reputation and the evaluated value of the comprehensive reputation work as input and output separately in the model. The comprehensive reputation $p = (p_1, p_2, \cdots, p_q)^T$ is hazed with symmetric triangular fuzzy number of $P_i(p_i, c_i)$, ($1 \le i \le q$). c_i is to be determined according to actual situation. The fuzzy number of comprehensive reputation of cross-domain users is represented with $P^*{}_i\left(a_0 + \sum_{j=1}^{S} R_{ji} a_j , e_0 + \sum_{j=1}^{S} R_{ji} e_j \right)$.

The fuzzy coefficient A_j is hazed with symmetric fuzzy number of $A_j\left(a_j, e_j\right)$. Since it is necessary to determine the fuzzy coefficients of $A_0, A_1, A_2, \cdots, A_S$ before determining the model, the question is converted into how to determine the coefficients of a_j and e_j (j=0…S). According to the fuzzy linear regression theory, the degree of fitting of comprehensive reputation is $h_i = 1 - \left| p_i - (a_0 + \sum_{j=1}^{S} R_{ji} a_j) \right| \bigg/ (c_i + e_0 + \sum_{j=1}^{S} \left| R_{ji} \right| e_j)$, when and only when $0 \le h_i \le 1$, the fitting degree h_i between the predicted value $P_i\left(p_i, c_i\right)$ and the evaluated value makes a sense, otherwise h_i is 0.

For a sampled value of $P_i\left(p_i, c_i\right)$ of a cross-domain user, when the central position $a_0 + \sum_{j=1}^{S} R_{ji} a_j$ of $P^*{}_i$ is fixed, the higher the fuzzy degree $S_{P_i^*}$ is, the higher h_i will be. So the optimal regression should make the total fuzzy degree of $A_0, A_1, A_2, \cdots, A_S$ reach minimum value. Set $S = \sum_{i=1}^{q} \omega_i S_i$ to represent the total fuzzy degree of the fuzzy coefficients. Then according to that h_i should not be less than desired value the restrictive conditions will be:

$$\begin{cases} \sum_{j=1}^{S} R_{ji} a_j + (1-H) \sum_{j=1}^{S} \left| R_{ji} \right| e_j \ge p_i - (1-H)\delta_i \\ \sum_{j=1}^{S} R_{ji} a_j - (1-H) \sum_{j=1}^{S} \left| R_{ji} \right| e_j \le p_i + (1-H)\delta_i \end{cases} \tag{3}$$

H represents the desired degree of fitting, e_j is positive and will be determined according to actual need. According to the definition of fuzzy degree, $S_i = \frac{1}{3}\left(e_0 + \sum_{j=1}^{S} R_{ji} e_j \right)$. So the target function will be simplified as $\min S = \left(\sum_{j=0}^{S} w_j e_j \right)$. There are two sets of the solutions to equation (3) according to whether a_j equals to 0.

3 Experiments and Analysis

3.1 Calculation of Comprehensive Reputation and Verification of Reliability

As shown in fig.1, three college LANs are three domains in the experiment. Totally there are four categories of applications, several of which each domain provides. The assessment standard of user's reputation in each kind of application may be different, but they all are recorded in form of score. Data of reputation scores in each domain sampled for 8 months are listed in table 1.

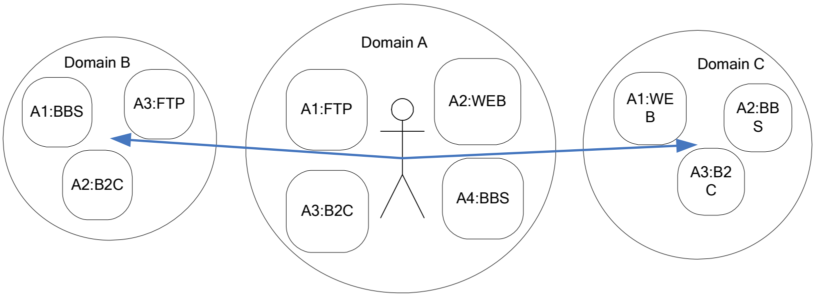


Fig. 1. This shows the topological diagram of domains in experiment. Among the three domains, domain A is the user's primary domain. There are several applications in each domain for users' visitation.

Table 1. Sampled user's reputation record of each application

	Domain A (32 times)				Domain B (15 times)				Domain C (13 times)			
	A1	*A2*	*A3*	*A4*	*A1*	*A2*	*A3*	*A4*	*A1*	*A2*	*A3*	*A4*
1	21	20	13	5	15	17	10	2	18	13	12	5
2	23	18	16	5	15	17	10	3	18	16	14	5
3	25	19	15	5	15	19	10	5	19	15	13	5
4	26	20	15	5	15	18	10	6	20	15	14	5
5	26	22	14	5	15	19	10	8	21	14	15	5
6	27	22	16	5	15	20	10	12	22	16	15	5
7	27	23	18	5	15	20	10	12	23	18	15	5
8	29	21	16	5	15	20	10	11	23	16	16	5

Firstly, weighting factors are determined according to visit times for each domain, and the reputation of cross-domain listed in table 2 is calculated out with data in table 1 according to equation 1. The last column in table 2 is the score marked by administrator of cross-domain. Secondly, fuzzy regression model can be built and data of table 2 is substituted into equation 3. Lastly, triangular fuzzy number is shown in table 3 figured out with the assistance of MATLAB.

As shown in table 4, to compare the original score with the regression value of P_i^*, the relative deviations are all less than 15%, the degrees of fitting are larger than 0.6, which is acceptable.

Table 2. Calculated reputation value of cross-domain

q	R_{1i}	R_{2i}	R_{3i}	R_{4i}	P_i (score)
1	18.85	17.73	12.03	4.25	5
2	19.92	17.32	14.07	4.50	5.5
3	21.20	18.13	13.32	5.00	6.1
4	21.95	18.42	13.53	5.25	6.2
5	22.17	19.52	13.22	5.75	6.2
6	22.92	20.20	14.28	6.75	7.3
7	23.13	21.17	15.35	6.75	7.6
8	24.20	19.67	14.50	6.50	6.5

Table 3. Calculated reputation of fuzzy number

j	Centre Value a_j	Fuzzy Amplitude e_j	Triangular Fuzzy Number A_j
	-12.32	0.03	(-12.32, 0.03)
	-0.31	0.08	(-0.31,0.08)
	0.77	1.06	(0.77,1.06)
	0. 05	0.02	(0.05,0.02)
	0.56	0.24	(0.56,0.24)

Table 4. Relative deviations calculated out

Time stage	1	2	3	4	5	6	7	8
Deviation	0.12	0.09	0.12	0.10	0.10	0.09	0.08	0.09
Degree of fitting	0.62	0.76	0.84	0.81	0.80	0.86	0.74	0.75

3.2 Average Precision Analysis

In the model system, frequency of users' visitation to every domain has obvious affection on users' reputation.

Definition 2. AVD is the short name of a user's average visiting density. M is sum of visiting times for all the applications in all the related domains (including both primary and active domains) of a certain user during considered time stage. N is the multiplication of sampling times in each domain and the number of applications in the same domain. If M>N, AVD is 100%, else AVD is equal to M/N.

To study how AVD affects the average precision of reputation calculation, 30 users are divided into 3 groups with 10 users a group according to their AVDs.

Figure 2 shows some characteristics obviously. In general, with improvement of the duration of sampling time, regression degree of reputation turns out to be an obvious increasing trend and more accuracy.

It's clear that regression degree changes according to AVD. This proves that AVD has obvious affection on relative error of the model. When AVD reaches 80% or higher level, regression degree gets to be more than 0.5. While AVD is only 40%, regression degree of comprehensive reputation is lower than 0.5 which is considered to be not

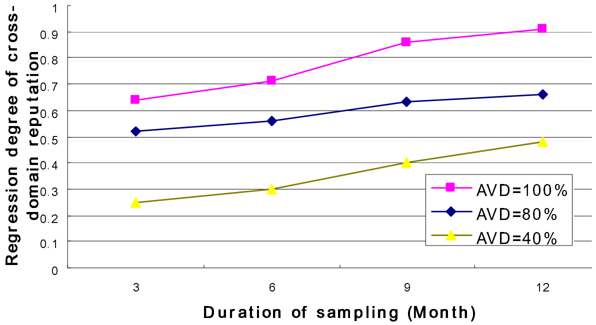


Fig. 2. This shows the comparison of regression degree under different AVDs

accurate enough. This phenomenon can be explained that if users visit certain application more frequently, their behavior comes near to their actual state. And this is consistent with the common-sense that cheating and disguise are usually sporadic. And when AVD is very low, much of sampling data should be filled up to complete the model. This can lead to larger error because the filled data maybe untrue. So the sampling interval should be selected properly and duration of sampling time should be long enough giving consideration to reputation update frequency for each application to avoid the above problem.

4 Conclusions

This paper brings out a comprehensive reputation calculation model based on fuzzy regression to provide uniform reputation evaluation method for cross-domain users and supply reputation sharing for corporation of cross-domain in network.

The following work includes analysis on the characteristics of reputation estimation of all categories of application, the way to improve the accuracy of reputation expression of cross-domain. And the store and management of cross-domain reputation data is also another issue to study.

Acknowledgments. The authors wish to thank all colleagues who previously provided technical support. The project is supported by “the Fundamental Research Funds for the Central Universities” (Grant No. 2010B23914).

References

1. CERNET, `http://www.cernet.edu.cn`
2. Feng, X., Jian, L.: Research and Development of Trust Management in Web Security. Journal of Software 13(11), 2057–2064 (2002)
3. Cui, Y.: Research on Trust Model and Access Control Model in Grid Environment. Dalian University of Technology, Dalian (2009)

4. Demchenko, Y., de Laat, C.: Domain Based Access Control Model for Distributed Collaborative Applications. In: Proceedings of the Second IEEE International. Conference on e-Science and Grid Computing, e-Science 2006 (2006)
5. Xiong, L., Liu, L.: A reputation-based trust model for peer-to-peer ecommerce communities. In: Proceedings of the IEEE Conference on E-commerce (2003)
6. Resnick, P., Zeckhauser, R., Friedman, R., Kuwabara, K.: Reputation Systems. Communications of the ACM 43(12), 45–48 (2000)
7. Zhang, S., He, D.: Fuzzy Model for Trust Evaluation. Journal of Southwest Jiaotong University 14(1), 23–28 (2006)

A TV Commercial Detection System

Yijun Li and Suhuai Luo

School of Design, Communication and Information Technology
The University of Newcastle, NSW2308 Australia
yijun.li@nielsen.com, Suhuai.Luo@newcastle.edu.au

Abstract. Automatic real-time recognition of TV commercials is an essential step for TV broadcast monitoring. It comprises of two basic tasks: rapid detection of known commercials that are stored in a database, and accurate recognition of unknown ones that appear for the first time in TV streaming. In this paper, we present the framework of a TV commercial detection system.

1 Introduction

TV Advertising plays an important role in our lives. Today, TV is the most effective media for marketing new products. Advertisers spend billions of dollars every year to produce advertisements on TV. CTR Market Research data show that in 2009 the Chinese media, advertising spend amounted to $74 billion, representing a 13.5% growth from 2008. TV advertising spend remained the predominant media, and its advertising income grew 15 percent, far ahead of any other media in 2009. In Australia, Nielsen Media Research estimates that advertising spend for 2009, across both metropolitan and regional television equates to 38% of total advertising or $3.5 billion dollars. There are many objectives for automatic TV commercial detections. Referring to the illustrative paradigm in Figure 1.1, we summarize four points to explain the motivations and potential applications of TV commercial video segmentation, indexing and identification. Firstly, many companies in marketing research and advertisement are interested in identifying commercial breaks from live TV streams. They may want to verify if a TV commercial has actually been broadcasted as contracted; and they may also want to know how their competitors are conducting their advertisements. It is highly desirable to have an efficient system for automatic recognition of TV commercials and storing the recognized advertisements in a video database, which can be retrieved based on video content or textual information from the database when requested. Secondly, many audiences do not like TV commercials. They may want to record and watch TV programs and exclude TV commercials. With the advancement of Personal Video Recording products (PVRs) in terms of large of storage, it is desirable to have a TV commercial skipping system, which detects and skips commercial automatically. Thirdly, all advertisements deal with one of three concepts: ideas, product, and services [27]. TV commercial classification with respect to the advertised products or services (e.g., automobile, finance, etc.) help to fulfill the commercial filtering towards personalized consumer services. For example, the MMS message (containing Key frames or adapted video) can send the commercials of interest to a registered user's mobile device or email box. Fourthly, the technology of TV commercial has changed a lot as they are almost

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 35–43, 2011.

always edited on a computer. The appearance all starts with MTV and MTV commercials are more visual, more quickly paced, use more camera movement, and often combine multiple looks, such as black and white color, or still quick cuts [27]. Accordingly, a TV commercial archive system including browse, classification, and search may inspire the creation of a good commercial. Marketing companies may even utilize it to observe competitor's behaviors [7].

TV Commercial Recognition
TV Advertisers: Is my product broadcasted as contracted?
TV consumers: Can I record TV programs by skipping TV commercials?
MMS and Email: Can I send Key frames of TV commercials to users by MMS or Email?
Market Research & Agency: Top 100 TV commercial last year.

Fig. 1.1. The applications of TV commercial Recognition

With the development of digitization technology, compression and archiving of multimedia Information, it is inexpensive to store Gigabytes of video in the computer for later retrieve. Basically, there are three ways of finding new or existing commercials from TV streams or Video database.

1. Manual editing. Users browse through recorded TV streams to identify the commercial break. They find individual commercials and query the TV station log to find commercial details.
2. Text-based retrieval. Textual information is added and used to guide conventional, text-base query.
3. Content-based retrieval. A content based search engine first identifies the commercial break and matches each individual commercial by using their audio\visual contents automatically.

The first two methods have several limitations. Firstly, they are all manual processes. It is inefficient and expensive because it needs considerable amount of human time and efforts. Secondly, it is imprecise because it is associates subjective human perception of the content being annotated. Due to the inefficiencies and limitations of the first two approaches, it is desirable to develop a system based solely on its content, which is the third approach. Detecting TV commercials from long video streams and retrieving similar or identical TV commercials from a centralized database by desktop graphic user interface (GUI) or World Wide Web (WWW) are the crucial problems that are highly related to video processing techniques, such as video segmentation, feature extraction, feature vector organization, indexing and retrieval. Automatic detection of commercials from TV broadcasting has attracted a lot of recent attention from both the research community and the marketing industry. Existing commercial detection approaches can generally be divided into two categories:

1. Feature-based approaches [14, 18, 19, 20] and
2. Recognition-based approaches [1, 7, 8, 12, 13, 15, 24].

While the feature-based approaches use some inherent characteristics of TV commercials to distinguish commercials and other types of videos, the recognition-based methods attempt to identify commercials by searching a database that contains known commercials. The challenges faced by both approaches are the same: how to accurately detect commercial breaks, each of which consists of a series of commercials; and how to automatically perform fast commercial recognition in real time.

A representative solution to automatic recognition of TV commercials is proposed by [20]. It utilizes monochrome frames and scene change ratio as pre-selectors, and each commercial break is determined by the edge change ratio and motion vector length. The main advantage of this approach comes from its high efficiency of detection, thus is appropriate for real-time commercial recognition. This type of feature-based approach can also be used as a filtering technique to accelerate commercial recognition (e.g., [17]). However, this approach incurs the following drawbacks. A generic threshold that is suitable for different TV channels and programs is very difficult to obtained, thus a detection system based on this approach is sensitive to TV broadcasting. Meanwhile, when black frames are not used at the beginnings and ends of some commercial breaks (this is common for some TV stations), this type of approach will fail. Another problem is that scene changes in commercials and some action movies can be very similar. In other words, this approach can miss some commercials that are not started with black frames, or falsely identify some fast-moving movies as commercials.

A problem parallel to the problem investigated in this paper is to automatically identify commercials from radio broadcasting. A recent piece of work proposes to use audio fingerprints to detect repeated objects in order to locate the commercial breaks [13]. The highly compact audio signatures are used to efficiently process the audio stream and audio-visual streams. With the audio information, the unknown commercials can be recognized from audio-visual streams in real time. However, this approach assumes that repeated objects contain both same visual and same audio information, thus it is highly dependent on the audio information. For videos without audio information or those with distorted audio information, this approach becomes unworkable. Duan et al. transformed the boundary detection of individual commercials into the problem of binary classification of shot boundaries by using the video frame information and the audio change information [7]. However, just as its previous work [13], accurate recognition of commercials still depends on the availability of audio information in video streams.

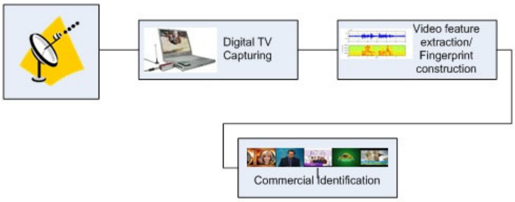

Digital TV Capturing
Video feature extraction/ Fingerprint construction
Commercial Identification

Fig. 1.2. TV Commercial Recognition System

A typical automatic TV commercial recognition system consists of the steps as shown in Figure 1.2. Firstly, live digital TV video data is captured and stored in a computer system. All commercial breaks are identified and segmented into commercial spots (i.e., individual commercials). There are several research efforts [2, 20, 18, 19, 22] addressing this non-trivial issue. Secondly, a fingerprint of each commercial spot is extracted. A commercial spot comprises a sequence of image frame. Thirdly, a new commercial spot appearing for the first time should be recognized by a human, and its fingerprints and identifier, such as the key number, are stored in the video database. The key number is an identity for a commercial spot provided from advertisers or marketing agency for commercial spot identification. Finally, a new commercial spot will be compared against commercial spots stored in the video database to check if it is similar to any known commercial spots, using their fingerprints. The last step is the most difficult one, in terms of both effective similarity matching and efficient query processing. Due to many factors introduced by different broadcasting stations, digitizing and recording methods, the same commercial video sequence could be different (for example, slightly different frame rate and colour variations). On the other hand, different commercial spots may have very similar video contents.

There are a number of efforts [4, 9, 23, 10, 25] in video sequence matching using video signatures. A standard method for retrieving video is to use content-based image retrieval (CBIR) techniques to seek frames with similar content. Considering colour features, there are two ways to extract colour-based features: by shot or by frame. The shot-based approach generally selects one or two representative frames from a shot (called key frames). A video sequence can be reduced to a small set of key-frames [21, 26, 29, 28, 5], and then CBIR techniques can be employed to match key frames [3]. These types of methods suffer from the fact that it is not clear as to which frame should be used for a shot, and the "action" (such as change of contents) within video sequences are largely ignored. The frame-based approach simply compares every individual frame between two clips in order to find sequences of frames that are consistently similar [2, 20, 18, 19, 22, 6]. The frame-based approach seems to be accurate, however, this type of signature will be very large and not efficient in indexing a large video database. The above observations motivate us to find a new approach for detecting video sequence similarity.

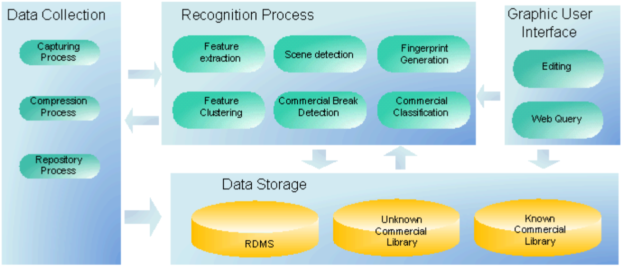


Fig. 2.1. System Architecture of TV commercial Detection System

2 TV Commercial Detection System Architecture

The TV commercial detection system is made of four major components – The collection process, the recognition process, data storage management and graphic user interface as illustrated in Figure 2.1. In the collection process, the system records broadcast TV data to the hard disk, and hand that recording data to the commercial break recognition process which automatically identifies each commercial break from recorded data. The system converts the record high quality TV data to a low quality TV and discards all recorded high quality material except for the commercial break. Once the commercial break has been identified they are sent to the recognition process. The recognition process loads the commercial break and compares them against the known commercials library. The system automatically classifies all advertisements which has been previously captured and recognized. An operator classifies the details of the new advertisement and makes it available for automatic recognition in the next time when it broadcasts.

A. Data Collection Process

The data collection process runs on a computer which has a digital TV tuner card (i.e. 16 card [16] or 11 card [11]). The process captures DVB television broadcast and encodes the programmes according to the MPEG-2 digital video standard with the audio signal coded in line with the Layer-III format. DVB creates the first standard for DVB-S (digital satellite TV), DVB-C (digital cable TV) and DVB-T (digital terrestrial TV). An example of HDTV video signal has 720 × 1280 pixels/frame, progressive scanning at 50 frames/ per second, requires.

$$\frac{720\ \times 1280\ \text{pixels}}{\text{frame}} \times \frac{50\ \text{frames}}{\text{second}} \times \frac{3\ \text{bytes}}{\text{pixel}} \approx \ {}^{138\ \text{megabytes}}\!/_{\text{per second}}$$

The HDTV format needs 250 gigabyte disk space roughly for recording to 30 minutes TV broadcast.

The raw data is compressed in MPEG-4 format with 3000 Kbit/s data rate, 25 frames /per second frame rate, and with a key frame every 8 seconds.

Each TV channel (station) that is captured requires 2 dedicated capture machines, one to be a “master” or primary one (also called a “hot” machine) and other serving as a “slave” or backup (also called a warm machine). Both machine would continuously capture TV broadcast, the only difference is that the hot machine would store (register) its data within the SQL database and warm one would not. In the instance that something goes wrong with the hot machine, data is automatically “patched” from the warm machine.

B. Recognition Process

The recognition process is running parallel with data collection process. Once raw data is received, it performs feature extraction step.

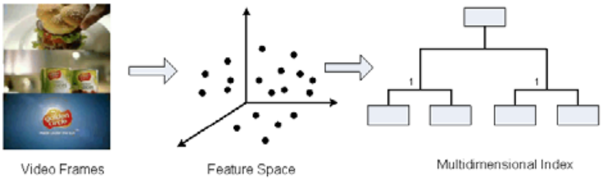


Fig. 2.2. Mapping video frames to multidimensional feature space

The visual information of frames can be expressed in the combination of text and computer graphics. The textual section may consistent of commercial brand name, advertiser name and contact details. Alongside the textual, a drawing or a photo of product might be placed with computer graphic technique. Basically, feature extraction maps from a large information space to a smaller feature space. As shown in Figure 2.2, feature extraction process reduces the complexity of the visual information and maps each video frame into a k-dimension feature space, and builds a multi-dimension index structure storing the feature space representation.

Generally, the features chosen to be extracted should follow some rules. The features should carry enough information about video frames and should not require any domain-specific known for their extraction. They should be small size and easy to compute in order to efficiently retrieve any interesting candidates in a large video database. They should relate well with human perceptual characteristics so that users can determine the suitability of the retrieved candidate video. They should be robustness to geometric and signal based distortion.

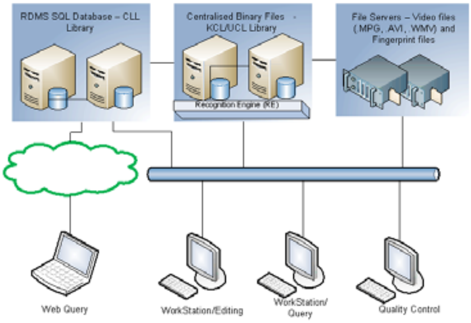

RDMS SQL Database – CLL Library
Centralised Binary Files - KCL/UCL Library
File Servers – Video files (.MPG, .AVI, .WMV) and Fingerprint files
Recognition Engine (RE)
Web Query
WorkStation/Editing
WorkStation/ Query
Quality Control

Fig. 2.3. The Architecture of Data Storage Process

C. Data Storage Process

The system contains three types of reference libraries - *Commercial Links Library* (*CLL), Known Commercial Library (KCL), and Unknown Commercial Library (UCL).* The CLL data are stored in a relational database such as Microsoft SQL Server and Oracle. CLL contains the details of a commercial such as commercial identification (ID), product information, duration of commercial, date and time of broadcasted, channel information etc. Those information are entered when either a new creative identified by the system or an old creative is matched automatically. The CLL library has an initial set of spots whose details are entered by an operator. KCL/UCL libraries are serialised binary files containing integer value of commercial IDs and fingerprint of individual commercial. KCL/UCL are centralised libraries located on the Server and are accessed by recognition process through Recognition Engine (RE). The Recognition Engine searches KCL/UCL library when a query is raised from the network. The Recognition Engine compares the fingerprint of a query against the fingerprints stored in the KCL/UCL library and returns a matched candidate whose ranking value is greater than a predefined threshold. The Recognition Engine is a multi-threads process and is configured as a singleton object handled all queries in the queue. The KCL library contains all known commercials that are validated by operators. The UCL library contains the new commercials that are broadcasted for the first time and have not been confirmed by operators. For example, assume that the CLL contains almost all broadcasted spots. So if a new commercial is broadcasted for the first time and is surrounded by known commercials, by detection the commercial break precisely, the system will learn a new commercial appeared. Thus, an unknown commercial is added into UCL library. The Figure 2.3 demonstrates the architecture of data storage process. In the system, a user can submit a query through a web or an application by providing a sample clip, a video name, a URL or a product name and specify the search mechanism he/she wants. By using the functionality of RE, the user can easily identify which mechanism returns more accurate results.

D. Graphic User Interface

The main Graphic User Interface (GUI) of the system is the TV editing screen as shown in Figure 2.4. The TV editing program runs on the operator workstation and is designed to verify auto matched commercial and identify new commercial. The TV editing operator views and monitors a 24-hour log of TV station broadcast consisting of all programs and advertisement broadcast from mid-night to mid-night. In the TV editing program the operator will identify any programs and new commercials or non-commercial activities.

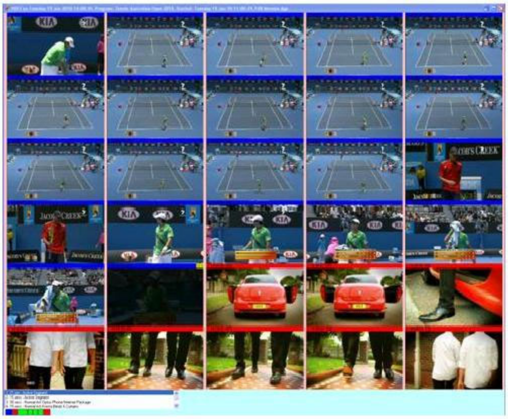


Fig. 2.4. Graphic User Interface

3 Conclusion

In this paper, commercial detection methods have been studied. The proposed TV commercial detection system is made of four components – data collection, recognition, data storage management and the GUI. In the collection process, the system record live digital TV content to the hard disk and hand that recording data to the commercial break recognition process which automatically identifies each commercial break from the recorded data. The recognition process loads the commercial break and compares them against the known commercial library. The system automatically classifies all advertisements which have been previously classified and recognize them.

References

1. Albiol, A., Fulla, M.J., Albiol, A., Torres, L.: Detection of TV Commercials. In: International Conference on Acoustics, Speech and Signal Processing, pp. 541–544 (2004)
2. Angihotri, L., Dimitrov, N., McGee, T., Jeannin, S., Schaffer, D., Nesvadba, J.: Evolvable Visual Commercial Detector. In: Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), vol. 2, pp. 79–84 (2003)
3. Ardizzone, E., et al.: Content-based indexing if image and video database by global and shape features. In: Proc. Of the International Conference on Pattern Recognition (1996)
4. Chang, C.L.E., Wang, J., Wiederhold, G.: Rime: A replicated image detector for the World Wide Web. In: SPIE Multimedia Storage and Archiving Systems III (1998)
5. Chang, H.S., Sull, S., Lee, S.U.: Efficient video indexing scheme for content-based retrieval. IEEE Trans. Circuits Syst. Video Technology (1999)
6. Cheung, S.-C.S., Zakhor, A.: Efficient video similarity measurement and search. In: Proc. of International Conference on Image Processing, British Columbia, Canada, pp. 85–89 (2000)
7. Duan, L.Y., Wang, J., Zheng, Y., Jin, J.S., Lu, H., Xu, C.: Segmentation, categorization, and identification of commercial clips from TV streams using multimodal analysis. In: ACM Multimedia, pp. 201–210 (2006)
8. Gauch, J.M., Shivadas, A.: Finding and identifying unknown commercials using repeated video sequence detection. In: Computer Vision and Image Understanding, vol. 103, pp. 80–88. Elsevier Science Inc., New York (2006)
9. Hampapur, A., Bolle, R.M.: Feature based indexing for media tracking. In: Proc. of Int. Conf. on Multimedia and Expo, pp. 67–70 (2000)
10. Hampapur, A., Hyun, K., Bolle, R.M.: Comparison of sequence matching techniques for video copy detection. In: Proc. SPIE, vol. 4676, pp. 194–201 (2002)
11. http://www.11.com/
12. Herley, C.: Accurate repeat finding and object skipping using fingerprints. In: ACM Multimedia, pp. 656–665 (2005)
13. Herley, C.: ARGOS: automatically extracting repeating objects from multimedia streams. IEEE Transactions on Multimedia 8, 115–129 (2006)
14. Hua, X., Lu, L., Zhang, H.: Robust Learning-Based TV Commercial Detection. In: ICME, pp. 149–152 (2005)
15. Hua, X., Chen, X., Zhang, H.: Robust Video Signature Based on Ordinal Measure. In: International Conference on Image Processing (ICIP), pp. 415–423 (1998)
16. http://www.16.com.au/
17. Kashino, K., Kurozumi, T., Murase, H.: A quick search method for audio and video signals based on histogram pruning. IEEE Transactions on Multimedia 5, 348–357 (2003)
18. Li, Y., Jin, J.S., Zhou, X.: Video Matching Using Binary Signature. In: Proceedings of the 2005 International Symposium on Intelligent Signal Processing and Communications Systems (ISPAC 2005), Hong Kong, December 13-16, pp. 317–320 (2005)
19. Li, Y., Jin, J.S., Zhou, X.: Matching Commercial Clips from TV Streams Using a Unique,Robust and Compact Signature. In: DICTA, Australia, pp. 266–272 (2005)
20. Lienhart, R., Kuhmunch, C., Effelsberg, W.: On the detection and recognition of television commercials. In: ICMCS, pp. 509–516. IEEE Computer Society, Los Alamitos (1997)
21. O'Connor, B.C.: Selecting key frames of moving image document: A digital environment for analysis and navigation. Microcomputers for Information Management 8(2), 119–133 (1991)

22. Nafeh, J.: Method and Apparatus for Classifying patterns of Television Programs and Commercials Based on Discerning of Broadcast Audio and Video Signal, US patent 5, 343, 251 (1994)
23. Sanchez, J.M., Binefa, X., Radeva, P.: Local colour analysis for scene break detection applied to tv commercial recognition. In: Proc. of Visual 1999, pp. 237–244 (1999)
24. Sanchez, J.M., Binefa, X., Vitria, J.: Shot Partitioning Based Recognition of TV Commercials. In: Multimedia Tools Applications, vol. 18, pp. 233–247. Kluwer Academic Publishers, Hingham (2002)
25. Shen, H.T., Ooi, B.C., Zhou, X.: Towards Effective Indexing for Very Large Video Sequence Database. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 730–741 (2005)
26. Tonomura, Y., Abe, S.: Content orient visual interface using video icons for visual database systems. Journal of Visual Languages and Computing' 1, 183–198 (1990)
27. Vilanilam, J.V., Varghese, A.K.: Advertising basics! A resource guide for beginners. Response Books, New Delhi (2004)
28. Wolf, W.: Key frame selection by motion analysis. In: Proc. ICASSP 1996, vol. II, pp. 1228–1231 (1996)
29. Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: Proc. IEEE ICIP 1998, vol. 1, pp. 866–870 (1998)

Research and Implementation of Entropy-Based Model to Evaluation of the Investment Efficiency of Grid Enterprise

Kehe Wu[1], Xiao Tu[1], and Cheng Duan[2]

[1] School of Control and Computer Engineering
[2] College of Electrical and Electronic Engineering
North China Electric Power University
wukehe@ncepu.edu.cn,
aiwotx@126.Com,
duancheng1985@hotmail.com

Abstract. Investment efficiency on power grid enterprises evaluation is an evaluation of proportional relationship between investment results and investment consumption. The rationality of evaluation depends on evaluation model, the most important part of evaluation model is the enactment of weight. As a long-tested weight method, entropy weight coefficients method is reliable and objective. In this article, we discussed a entropy weight coefficients method that applicable to the investment efficiency on power grid enterprises evaluation. The computer realization method is also disscussed in this article.

Keywords: entropy weight coefficients method, J2EE, power grid, investment efficiency.

1 Introduction

With the continuous development of the national economy and increasing living standards, the social demand for electricity and the power grid annual investment is also increasing. In order to avoid ineffective investment, arrange investment in science, it is necessary to build a scientific and effective evaluation model to access the investment result on power grid. To build evaluation model, index system should be established first, then set the index weights of the index system. Index weights reflects the importance of the index in the accessment. Different weights have a direct impact on the evaluation results. So it is critical important to choose a good index weights method.

Index weights method is a method that define index weights by judgment matrix consist of index value in objective condition. It try to eliminate the subjectivity of the weight to make assessment results more practical[1].

2 Index System of Investment Efficiency on Power Grid

The development of power grid is a complex project, the evaluation of the development is also a complex project with many attributes. To find a scientific and

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 44–50, 2011.

objective evaluation method to build a quantized evaluation index system is important. Each index in index system reflects different aspect of the investment efficiency situation. A reasonable index system not only can access the investment result in every aspect but also fairness. The value orientation of the power grid reflect in five aspects: The primary goal of the development of power grid is running secure and reliable, the ultimate goal is providing reliable and high-quality power to user, the higher goal is the sustainable development of power grid, the main aspect is the high-efficiency development of power grid, the trend of future development is Strong and Smart Grid. Considered from the characteristic and value orientation of power grid, the evaluation can be divided into five subsystems[2]. They are: safety, reliable, high quality, coordination and economic. Then build the evaluation index system by choosing index from these five points of views. And the indexs are: substation-installed capacity ratio, capacity-load ratio, N-1 checking, CO2 emission reductions, URT distribution, RSI, rural electrification level, wire maximum load factor, electricity sales amount, line loss per unit.

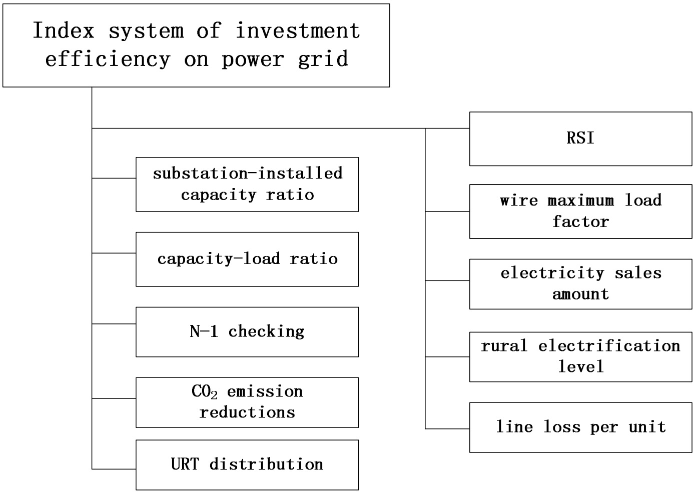


Fig. 1. Index system of investment efficiency on power grid

3 Evaluation Model Based on Entropy Weight Coefficients Method

The concept of entropy was originally derived from thermodynamics, to describe an irreversible phenomenon in motion. It was first introduced in information theory by Shannon, and widely used in engineering, socio-economic and other fields. In information theory, entropy is a measure of disorder level of the system. It is defined as follows: when the system can be in n different states, the probability of each state is $p_i(i=1, \ldots, n)$, then the entropy of the system is

$$E = -\sum_{i=1}^{n} p_i \ln p_i \tag{1}$$

In particular, when p_i values are equal, that is p_i=1/n, then the entropy has the maximum value E_{max}=lnn. Entropy reflects the degree of variation. The smaller the entropy of an index is, the bigger the degree of variation is. The bigger the degree of variation is, the more information it takes, and the more important it is in the evaluation, then the bigger the weight is. Otherwise, if the entropy of an index is big, it reflects that the degree of variation is small, so the weight is small. [3]Therefore, in the specific analysis process, we can according to the degree of variation of each index, using entropy to calculate the weight of each index, then weighting each index to get more objective evaluation results.

Entropy weight coefficients method can be divided into the following steps:

1) Assuming an evaluation in m indexs to n region samples, x_{ij} is the value of index j(j<=m) to sample i(i<=n), they consist the original matrix $X=(x_{ij})_{n*m}$.
2) In order to reflect the actual situation, to exclude the influence caused by unit or magnitude of different index and avoid unreasonable phenomenon, extremum method was used to make the indexs dimensionless in this article.
   a) Positive index is a kind of index that its value is the bigger the better, its dimensionless method is:

$$d_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, x_{\max} = \max(x_1, .. x_n), x_{\min} = \min(x_1, .. x_n) \tag{2}$$

   b) Negative index is a kind of index that its value is the smaller the better, its dimensionless method is:

$$d_i = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}}, x_{\max} = \max(x_1, .. x_n), x_{\min} = \min(x_1, .. x_n) \tag{3}$$

   c) Appropriate index is a kind of index that its value is better in an appropriate intevel, its dimensionless method is:

$$\begin{cases} d_i = \dfrac{x_{\max} - x_i}{x_{\max} - x_{up}}, d_i > x_{up} \\ d_i = 1, x_{low} < d_i < x_{up} \\ d_i = \dfrac{x_i - x_{\min}}{x_{low} - x_{\min}}, d_i < x_{low} \end{cases} \quad \begin{matrix} x_{\max} = \max(x_1, x_2 .. x_n), \\ x_{\min} = \min(x_1, x_2 .. x_n) \end{matrix} \tag{4}$$

   The x_{up} and x_{low} represent the upper limit and lower limit of the appropriate intevel.
3) According to the definition of entropy, the entropy of index j is:

$$E_j = -\frac{1}{\ln m}(\sum_{i=1}^{m} f_{ij} \ln f_{ij})(i = 1...m; j = 1...n)\ , f_{ij} = \frac{d_{ij}}{\sum_{j=1}^{n} d_{ij}} \tag{5}$$

To make sure formula is significant, when f_{ij}=0, we set $f_{ij}\ln f_{ij}$=0.

4) The weight of each index can be computed then:

$$W_j = \frac{1-E_j}{n-\sum_{j=1}^{n} E_j} \tag{6}$$

n is the amount of indexs. Obviously the sum of index weights is 1.

5) At last, the score of each region can be computed:

$$S_i = \sum_{j=1}^{n} d_{ij} W_j \tag{7}$$

For the indexs are in different dimensions, the dimensionless values are used instead of original values. The range of the dimensionless values are 0 to 1, so the scores are in 0 to 1.

4 The Implement of the Model

4.1 Application Designer Framework

The evaluation method described in this article is highly generability. So the extensibility of the system should be token full account. As a fully-fledged architecture, J2EE is high availability, high reliability and extensibility. The system will follow J2EE architecture, development using B/S architecture[4].

- Client Layer: Running on the client-side. It is used for showing the pages sending from the presentation layer, responding to user actions and sending messages to presentation layer. It is usually IE Browser in B/S architecture.
- Presentation Layer: Running on J2EE server. It is used for producing pages to user and handling the client's request, then sent to the business logic layer for processing. Struts framework is used in this system. Struts is a good combination of Spring Framework. If you add configuration of Spring plug-in in the Struts configuration file, you can initialize Spring during the initialization of Struts, and handle the action of Struts by Spring. Struts framework can divided into three parts, they are model, view and controller, and make sure every two parts of them are loose coupling. Structs also enrich JSP tags in view layer by customizing tags. It makes the web design more convenient [5].
- Business layer: Also running on the J2EE server. It is used for handling the business logic and managing transactions. Spring framework is used in the system. The core of Spring is IoC(Inversion of control)/DI(Dependence Injection). IoC is to extract the dependence of modules, and leave it to container to do the configuration. DI is a more vivid explanation to IoC. Its definition is to inject the dependence (for example, construction parameters, construction object or interface) dynamic to modules during the running time. There are three implement types in IoC/DI mechanisms. They are: interface injection (Factory pattern), constructor Injection(implement dependence in construction method), setter injection(implement dependence by setter method).

- Data persistence layer: Data persistence layer encapsulate the data access detail, and provide object-oriented API for business layer. Hibernate is used in this system to provide data access services. Hibernate is a fully-fledged O/R Mapping framework that supporting most of main current database. [6] It also support Parent/Child relation, transaction processing, extends, and polymorphism. The system use DAO (Data Access Object) design pattern in data persistence layer. By using DAO, data persistence layer encapsulates and abstracts all the accessing to data source and reduce the degree of coupling between business layer and itself. DAO completely hiding the implement detail of data source to client. Database is accessed by DAO when the data is needed. While the implement of data source is changing, the interface to client provide by DAO will not change, it will not affect client or business module.
- Database layer: Database layer used for saving basic data for business process. Oracle is used in this system.

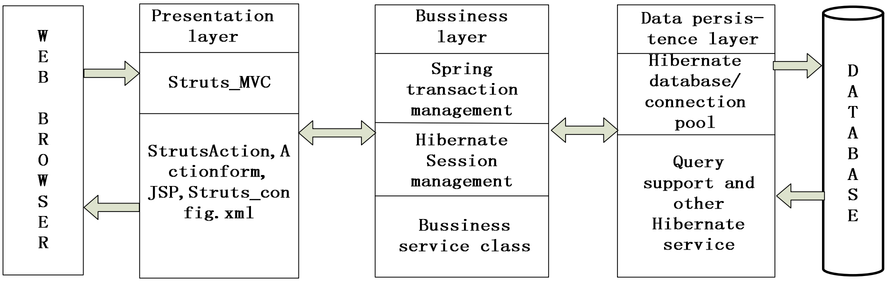

WEB BROWSER
Presentation layer
Struts_MVC
StrutsAction, Actionform, JSP, Struts_config. xml
Bussiness layer
Spring transaction management
Hibernate Session management
Bussiness service class
Data persistence layer
Hibernate database/ connection pool
Query support and other Hibernate service
DATABASE

Fig. 2. Web Application Framework

4.2 Implement of Program Function

1) implement of presentation layer: Presentation layer is used for showing index weights and index score to users. Jsp is used in this system. For a more personalized interaction, the system use fusioncharts for generating pie chart and column chart to showing the results. The correlative code is:

```
//transmit data for the pie chart
chart. setDataXML("${piexml}");
//define a column chart
var columnchart = new
FusionCharts("${basePath}resource/flashChart/Column3D.
swf", "ChartId2", width, height , "0", "0");
//transmit data for the column chart
chart. setDataXML("${columnxml}");
</script>
//quotation of fusioncharts
<script type="text/javascript
src="${basePath}js/FusionCharts. js"></script>
<script>
//define a pie chart
```

```
var piechart = new
FusionCharts("${basePath}resource/flashChart/Pie3D.
swf", "ChartId", width, height , "0", "0");
```

2) Implement of business layer:Business layer implement the business logic of evaluation calculation, including the business logic of weight calculation and score calculation.

```
public interface EntropyWeigh{
// dimensionless method•
public double[][][] toOne(double[][][] idxData){
}
// entropy weight coefficients method
public double[] entropyMethod(double[][][] idxData){
}
//evaluation method
public double[] evaluation(double[] weights, double[][]
idxData){
}
```

3) Data persistence layer: Data persistence layer used for handling data accessing operation. The Schema file is:

```
<?xml version="1. 0"?>
<!DOCTYPE hibernate-mapping PUBLIC "-
//Hibernate/Hibernate Mapping
DTD//EN" "http://hibernate. sourceforge. net/hibernate-
mapping-2. 0. dtd">
<hibernate-mapping>
<class name="IndexInfo" table="INDEXINFO">
   <id name="id" column="INDEX_ID">
      <generator class="increment"/>
   </id>
   <property name="name"
column="INDEX_NAME"  type="java. lang. String"/>
   <property name="idxvalue" column="INDEX_VALUE"
type="java. lang. Doule"/>
</class>
</hibernate-mapping>
```

5 Conclusion

In this article, entropy was introduced into the calculation of index weights, and build an evaluation model of investment Efficiency on power grid enterprises. As a strong objective evaluation method, entropy weight coefficients method reduces as much as possible human-subjective influence to the evaluation. Ensure the objectivity and reliability of evaluation results.

Index Weight Calculation Model is based on mathematical theory, calculation of complex weights by compute can help reducing the workload of manual, improve the accuracy of calculation, avoid human error resulting from the calculation. By using pleny of charts, the user experience is greatly improved. The program provides service to the Power Grid enterprise evaluation model, simple operation, flexible interface design, good usability. As a full-fledged weight method, entropy weight coefficients method is scientific and reliable.

References

1. Zheng, X.-h., Zhang, Q., Luo, m.: The Application of Entropy-Weight Coeffcient Method to Risk Decision. College of Economic Management, Wuhan Uni. (2000)
2. Yang, J., Zhang, X.-j., Sheng, H.-h., Lu, G.-l.: Research on Benchmarking Index Evaluation Method in Electric Power Enterprises. North China Electric Power (2009)
3. Fan, R., Wang, Z.: A Method of Entropy Weighting Ideal Point and Its Application in Investment Decision. Journal OF Wuhan University of Hydraulic and Electric Engineering (1998)
4. Lai, M., Wang, G.-s.: The Investigations and Calculations on the Weight Coefficients of the Post Evaluation Indexes for the Scientific and Technological Projects in Electric Enterprises. Central China Electric Power
5. Zhao, J., Wang, T., Niu, D.-x.: Improved entropy TOPSIS of knight service evaluation in electric power marketing. Journal of North China Electric Power University
6. Wang, X.-w., Meng, X.-s., Wang, F.-s.: The developing method of Web system based on SSH framework. Journal of Agricultural University of Hebei (Agriculture and Forestry Education Edition)

Passive Data Storage Based Housewares Store Management System

Yang Xiao, Guoqi Li, and Juan Zhang

School of Reliability and System Engineering, Beihang University,
Beijing, China
hwgl@hotmail.com, gqli@buaa.edu.cn,
zhangjuan198804@163.com

Abstract. After referring several cases of management information system and RFID system, we design a novel and practical system for housewares store. We use the RFID tags in a different way - Passive Data Storage. In the paper, the structure and implementation of the system is described in detail. A case study is also given for illustration.

Keywords: Management Information System, RFID, housewares store, passive data storage.

1 Introduction

With the development of information industry, people need to handle more data and information, so the management information system comes out. A management information System (MIS) [1] is a system that provides information needed to manage organizations effectively. It involves three primary resources: technology, information, and people. Thus, there is a typical problem that how to organize the three parts to work more effective.

An MIS is a planned system of the collection, processing, storage and dissemination of data in the form of information needed to carry out the management functions. The traditional systems need people to do more work for typing information. Even though, the technology has taken some conveniences, for example, powerful software can process information perfectly, it is seen that the three parts are also separate. We imagine a mode that information is to be related to the things, so that people just use technology to read the information from the target. Therefore, the processes of management become easier.

After many explorations, we finally focus on the RFID tags (Radio-frequency identification) [2]. We use the RFID tags in a different way. In some traditional occasion, the RFID tags just use for a mark, they just provide a number so that the supermarket or warehouse can identify them. But if we do not have the database of the tags, we can know nothing about the targets.

So, for more effective operation, we design a system for this management theory. The service object of the management system is housewares store. After referring several cases of management information system [3] [4] and RFID system [5], we design a rapid and practical system for housewares store. It mainly includes RFID,

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 51–56, 2011.

Bluetooth, and Embedded Software and other technologies. This management information system will show these advantages as follow:

- The information of housewares could be input to RFID by laptop and PDA;
- The RFID tags contain various information about the items;
- Paperless office;
- We use device around us just like laptop and PDA so that process will convenient and effective.

2 System Description

The Housewares Store Management System has four parts, can be seen in Fig.1, which include some passive data storage equipment, a handheld RFID reader/writer, a handheld personal digital terminal or a laptop.

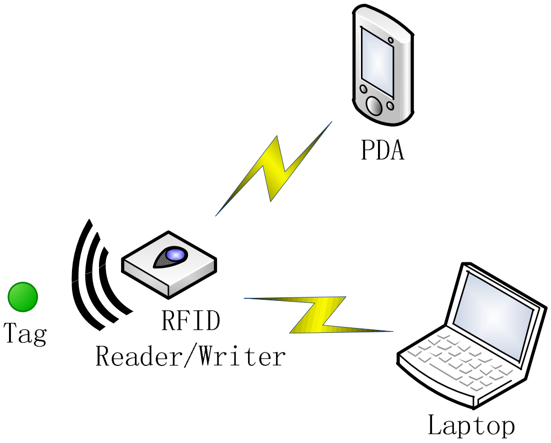


Fig. 1. Devices and their structure of the system

The passive data storage equipment, which is also called passive tag, is used for recording the information. And the handheld RFID reader/writer (we call it reader/writer for short) connects the tag with RF-Signal (Radio Frequency Signal). The handheld personal digital terminal, which is a PDA (Personal Digital Assistant) in our actual system, connects the reader/writer with Bluetooth signal. And the Laptop is also connected to reader/writer by Bluetooth signal.

The tag contain at least two parts: one is an integrated circuit for storing and processing information, modulating and demodulating a radio-frequency (RF) signal, and other specialized functions; the other is an antenna for receiving and transmitting the signal.

The reader/writer is composed of RFID module, Controller, Bluetooth module and Supply Circuit. The RFID module includes antenna, radio-frequency identification unit, and digital signal processing (DSP) unit. The RFID tag returns slight signal to

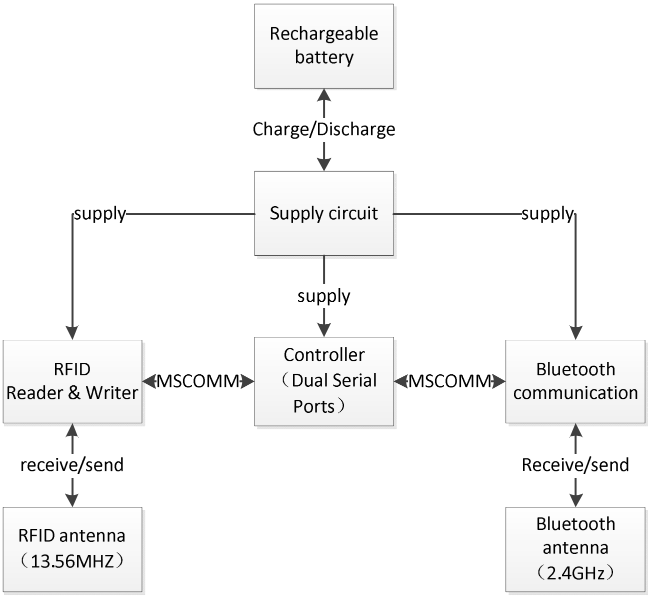


Fig. 2. The composition of RFID reader/writer

the antenna, then the radio-frequency identification unit turn the analog signal into digital signal. The DSP unit processes the signal, demodulate the information, and send the information to the controller with serial port. The controller transmits the information to the Bluetooth module. Then the Bluetooth module sends Bluetooth signal to the laptop or the PDA, so that we can require information from terminal. Similarly, we can write some information into the RFID tag by using the contrary path (The whole process can be seen in Fig.2).

The program in the terminal must be designed for completing the function of checking tag and writing information. Nowadays, the smart devices have been more and more popular. One of the conveniences for the smart device is the intelligent operating system. When we design the program, we can only focus on how to build the connection between devices. And we do not need to know how the connection works. The program sends and receives message through the serial port. We make a protocol for communication. The process of building connection is very simple. First, initialize the program, create two threads. One is to control the RFID reader/writer to have right operation; the other is to build the connection between the PDA and the RFID reader/writer, and maintain the connection. When the user has some operation, the program translates the operation into a command by the protocol, and sends the information to the RFID reader/writer. Then the program waits for the return value to show the result.

3 Software Implementation

The software implement parts include two types. One is running on the reader/writer, the other is running on the PDA.

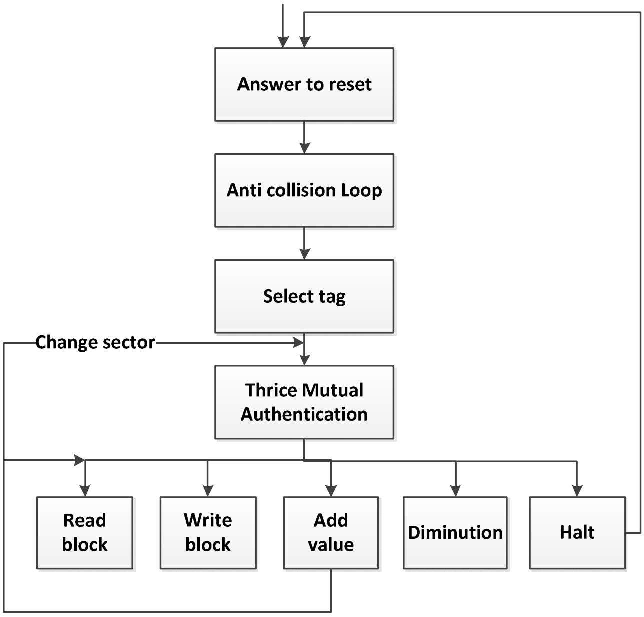


Fig. 3. The process of reading/writing tags

The program runs on the reader/writer is written by C-Language for MCU. When the RFID tag comes into the identification range of the RFID reader/writer, the reader/writer turns into work mode. The program must finish these missions. First, the reader/writer answers to reset, and the tag and the reader/writer must have same protocol and same communication baud rate. Then, there is an anti-collision loop for sometimes when several tags come into the identification range. This anti-collision loop will choose one tag to operate. After selecting tag, the reader/writer makes thrice mutual authentication to confirm the sector which to be accessed. The reader/writer checks the codes and then operates the data block on the tag. When the tag leave out of the range of reading, the reader/writer will halt, and turn to the initial state to wait another tag.

The program in the PDA terminal must be designed to complete the function of checking tag and writing information. Nowadays, the smart devices have been more and more popular. One of the conveniences for the smart device is the intelligent operating system. Here we use Windows CE as the operation system. Therefore, the message mechanism is mature [6]. When we design the program, we can only focus on how to build the connection between devices. And we do not need to know how the connection works. The program sends and receives message through the serial port. We make a protocol for communication.

The process of building connection is very simple. First, initialize the program, create two threads. One is to control the RFID reader/writer to have right operation, we call it "main thread"; the other is to build the connection between the PDA and the RFID reader/writer, and maintain the connection, we call it "communication thread".

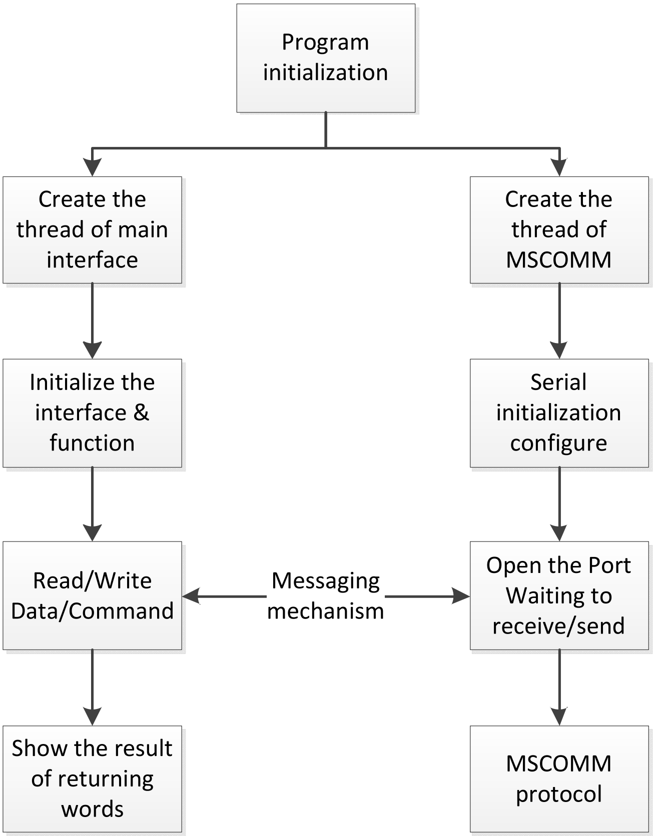

Program initialization
Create the thread of main interface
Create the thread of MSCOMM
Initialize the interface & function
Serial initialization configure
Read/Write Data/Command
Messaging mechanism
Open the Port Waiting to receive/send
Show the result of returning words
MSCOMM protocol

Fig. 4. The process of the software on PDA

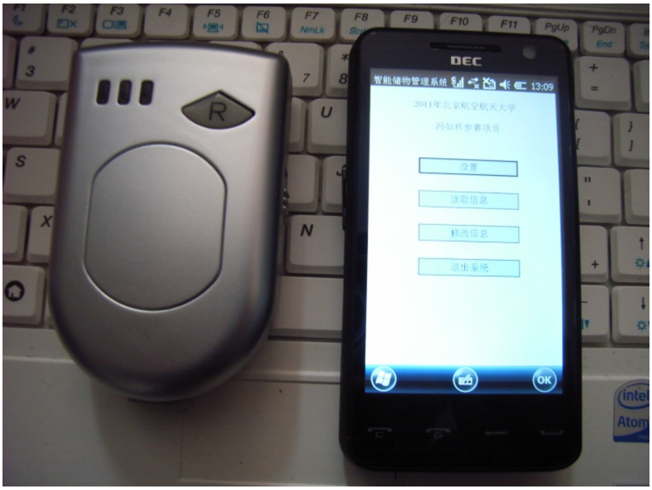


Fig. 5. Demonstration of the system

The main thread create operation interface. Meanwhile, the communication thread configures the port's parameters. When the user has some operation, the program translates the operation into a command by the protocol, and sends the information to the RFID reader/writer. Then the program waits for the return value to show the result.

The Fig.5 shows the devices that use in the management information system. The left side device is reader/writer; the right side device is a smartphone which uses Windows CE operation. We can see the operation interface on the screen.

4 Conclusion

We have made this system on trial for the housewares store. There is a good effect gained in practice. The combinatorial innovation always brings a new effective way to solve problem. The RFID technology is introduced into the management information system. We also consider the fastest method to achieve the goal, which reflects system engineering idea. The Passive Data Storage Based Housewares Store Management System gives people a more effective mode for management information system application. We use the RFID tag as another form of storage and identification. A different usage of the existing technology may lead a more effective solution. Therefore, if this system can be used in the similar condition, there may be a more amazing effect.

Acknowledgments. The technique and design of the system mentioned in this paper have been applied for a patent, Patent Application Number: 201120109605.6, 201110093831.4.

References

1. Serrano, N., Alonso, F., Sarriegi, J.M., et al.: A new undo function for Web-based management information systems. IEEE Internet Computing 9(2), _7 (2005), doi:10.1109/MIC.2005.28
2. Mo, J.P.T., Sheng, Q.Z., Xue, L., et al.: RFID Infrastructure Design: A Case Study of Two Australian RFID Projects. IEEE Internet Computing 13(1), _8 (2009), doi:10.1109/MIC.2009.18
3. Cao, X.H., Wan, J.A.: A RFID-Based Monitoring System for Abnormal Logistics Events in Internal Material Supply Chain Materials Science And Engineering. PTS 1-2 179-180 949-954 Part 1, 2 (2011)
4. Wang, L.C.: A RFID based agile manufacturing planning and control system. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS, vol. 5589, pp. 441–451. Springer, Heidelberg (2009)
5. Jolayemi, J.K.: A deterministic model for planning production quantities in a multi-plant, multi-warehouse environment with extensible capacities. International Journal of Production Economics 87, 99 (2004), doi:10.1016/S0925-5273(03)00095-1
6. Microsoft website, `http://msdn.microsoft.com/`

Multiple Solutions for Resonant Difference Equations[*]

Shuli Wang and Jianming Zhang[**]

Department of Mathematics, Taiyuan University of Technology, Taiyuan
Shanxi, 030024, P.R. China
{zwangshuli,tyutzjm}@163.com

Abstract. In this paper, the critical point theory, the minimax methods and Morse theory are employed to discuss the existence of nontrivial solutions for boundary value problems of second-order difference equations with resonance both at infinity and at zero. Some existence results are obtained.

Keywords: Difference equation, Resonance, Critical group, Morse theory, Local linking.

1 Introduction and Main Results

In this paper, we consider the following boundary value problems of second-order difference equations

$$(P)\begin{cases}-\Delta^2 u(t-1)=g(t,u(t)), & t\in[1,T],\\ u(0)=0,\ \ \Delta u(T)=0,\end{cases}$$

where $[1,T]:=\{1,2,\cdots,T\}$, $\Delta u(t):=u(t+1)-u(t)$ is the forward difference operator, $\Delta^2 u(t):=\Delta(\Delta u(t))$, with $g(t,0)=0$ for $t\in[1,T]$. Clearly, (*P*) has the trivial solution $u=0$. We are interested in finding nontrivial solutions for (*P*). The existence of nontrivial solutions of (*P*) depends on the local properties of *g* near infinity and near zero. Here, we consider the cases in which *g* is resonant both at infinity and at zero, i.e.*g* satisfies

$$g(t,x)=\lambda_k x+o(|x|),\quad \forall t\in[1,T],\ |x|\to\infty, \tag{1.1}$$

$$g(t,x)=\lambda_m x+o(x),\quad \forall t\in[1,T],\ x\to 0, \tag{1.2}$$

where λ_k,λ_m are two eigenvalues of the linear boundary value problem

$$(P_0)\begin{cases}-\Delta^2 u(t-1)=\lambda u(t), & t\in[1,T],\\ u(0)=0,\ \ \Delta u(T)=0,\end{cases}$$

The existence of solutions for difference equations has been studied by many authors, and some results were obtained by using different methods such as fixed point

[*] The project is supported by the National Science Foundation of Shanxi (No. 2011011002--4).

[**] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 57–64, 2011.

theorems or upper and lower solutions methods or critical point theory. For the details, we refer to $[3]-[5],[13],[15],[16]$ and the references therein. In this paper, we consider the existence of multiple solutions for (P) with resonance at both infinity and zero. We make the following assumptions:

$(f^{\pm})$ If $\frac{\|v_n\|}{\|u_n\|} \to 1$ as $\|u_n\| \to \infty$, then there exist $\varepsilon > 0, N \in \mathbb{N}$ such that

$$\pm \sum_{t=1}^{T} f(t, u_n(t)) v_n(t) \geq \varepsilon, \quad n \geq N,$$

where $f(t,x) = g(t,x) - \lambda_k x, u_n = v_n + w_n, v_n \in V, w_n \in V^{\perp}$.

$(g^{\pm})$ There exist $\delta, c_1, c_2 > 0$ and $\tau > 1$ such that

$$c_1 |x|^{\tau} \leq |g(t,x) - \lambda_m x| \leq c_2 |x|^{\tau},$$
$$\pm(g(t,x) - \lambda_m x)x \geq 0, \quad \forall t \in [1,T], \quad |x| \leq \delta.$$

The main results in this paper are the following theorems.

Theorem 1.1. Let (1.1) hold, and $g_x'(t,0) < \lambda_1$. Then (P) has at least four nontrivial solutions, among which one is positive and one is negative, provided that either of the following conditions is fulfilled:

(i) (f^{+}) and $k \in [2,T]$;

(ii) (f^{+}) and $k \in [3,T]$.

Theorem 1.2. Let (1.1) hold and $k = 1$. Then (P) has at least two nontrivial solutions in each of the following cases:

(i) $(f^{+}), (g^{+})$ and $m \in [2,T]$;

(ii) $(f^{+}), (g^{-})$ and $m \in [1,T], m \neq 2$;

(iii) $(f^{-}), (g^{+})$ and $m \in [1,T]$;

(iv) $(f^{-}), (g^{-})$ and $m \in [2,T]$.

2 Preliminaries

Let E be a Hilbert space and $J \in C^2(E,\mathbb{R})$ be functional possessing the deformation properties. Let u_0 be an isolated critical point of J with $J(u_0) = c$, and U be a neighborhood of u_0, containing the unique critical point, the group

$$C_q(J,u_0) := H_q(J^c \cap U, (J^c \setminus \{u_0\}) \cap U), \quad q \in \mathbb{N}$$

is called the q-th critical group of J at u_0, where $J^c = \{u \in E \mid J(u) \leq c\}, H_q(A,B)$ denotes the q-th singular relative homology group of the topological pair (A,B) with integer coefficients.

Denote $K=\{u\in E \mid J'(u)=0\}$. Assume that K is a finite set. Take $a<\inf J(K)$. The critical groups of J at infinity are defined by [2]

$$C_q(J,\infty):=H_q(E,J^a),\quad q\in\mathbb{N}.$$

Using these concepts, we have the following famous Morse inequality [7]:

$$\sum_{j=0}^{q}(-1)^{q-j}M_j\geq\sum_{j=0}^{q}(-1)^{q-j}\beta_j,\quad q\in\mathbb{N} \tag{2.3}$$

$$\sum_{q=0}^{\infty}(-1)^{q}M_q=\sum_{q=0}^{\infty}(-1)^{q}\beta_q. \tag{2.4}$$

where $E_k=\sum_{u\in K} rankC_k(J,u),\ \beta_k=rankC_k(J,\infty)$.

Let $u\in K$ be an isolated critical point of J such that $J''(u)$ is a Fredholm operator and Morse index $\mu(u)$ and the nullity $\nu(u)$ of u are finite. We have the following facts about the critical groups of J at u:

(i) $C_q(J,u)\cong 0,\quad q\notin[\mu(u),\mu(u)+\nu(u)]$.

(ii) If u is nondegenerate, i.e. $\nu(u)=0$, then $C_q(J,u)\cong\delta_{q,\mu(u)}\mathbb{Z}$.

Proposition 2.1. [12] Let 0 be an isolated critical point of $J\in C^2(E,\mathbb{R})$. Assume that J has a local linking at 0 with respect to $E=E^-\oplus E^+, l=\dim E^-<\infty$, i.e. there exists $\rho>0$ such that

$J(u)\leq 0,$ for $u\in E^-, \|u\|\leq\rho,$

$J(u)>0,$ for $u\in E^+, 0<\|u\|\leq\rho.$

Then $C_q(J,0)\cong\delta_{q,l}\mathbb{Z},$ if $l=\mu(0)$ or $l=\mu(0)+\nu(0)$.

Proposition 2.2. [2] Let the functional $J:E\to\mathbb{R}$ be of the form

$$J(u)=\frac{1}{2}\langle Au,u\rangle+Q(u), \tag{2.5}$$

where $A:E\to E$ is a self-adjoint linear operator such that 0 is isolated in $\sigma(A)$, the spectrum of A. Assume that $Q\in C^1(E,\mathbb{R})$ satisfies

$$\|Q'(u)\|=o(\|u\|),\ \|u\|\to\infty. \tag{2.6}$$

Write $V:=\ker A, W:=V^{\perp}=W^-\oplus W^+$ where $W^{\pm}$ are subspaces on which A is positive (negative) definite. Assume that $\mu=\dim W^-$ and $\nu=\dim V\neq 0$ are finite and J satisfies the deformation condition. Then

$$C_q(J,\infty)\cong\delta_{q,k^{\pm}}\mathbb{Z},\quad k^+=\mu,\quad k^-=\mu+\nu,\quad q\in\mathbb{N},$$

provided J satisfies the angle conditions at infinity:

$(AC_\infty^\pm)$ There exist $M>0$ and $\alpha\in(0,1)$ such that

$$\pm\langle J'(u),v\rangle\geq 0$$

for $u=v+w\in E=V\oplus W$, $\|u\|\geq M$, $\|w\|\leq\alpha\|u\|$.

3 Proofs of Main Results

Let $E=\{u:[0,T+1]\to\mathbb{R}\mid u(0)=0,\ \Delta u(T)=0\}$ denote a finite-dimensional real Hilbert space with inner product $\langle u,v\rangle=\sum_{t=1}^{T}u(t)v(t)$ and norm $\|u\|=\sqrt{\langle u,u\rangle}$. For any $u\in E$, denote $\|u\|_p=(\sum_{t=1}^{T}|u(t)|^p)^{1/p}$, $p\geq 1$. Then there exist $a_p,b_p>0$ such that

$$a_p\|u\|\leq\|u\|_p\leq b_p\|u\|,\quad\forall u\in E. \tag{3.7}$$

It follows from [9] that $\lambda_t=4\sin^2\frac{(2t-1)\pi}{4T+2}$, $t\in[1,T]$ are the distinct eigenvalues of (P_0), ϕ_t, $t\in[1,T]$ are the corresponding orthogonal eigenvectors, where $\phi_t(i)=\sin\frac{t(2i-1)\pi}{2T+1}$, $i\in[1,T]$. Then $\varphi_1>0$. For $k\in[2,T-1]$, E can be split as $E=W^-\oplus V\oplus W^+$, where

$$W^-=span\{\varphi_1,\cdots,\varphi_{k+1}\},\ V=span\{\varphi_k\},W^+=span\{\varphi_{k+1},\cdots,\varphi_T\}.$$

Every vector $u\in E$ can be written as $u=u^-+v+u^+$, where $u^-\in W^-$, $u^+\in W^+$, $v\in V$. Define the functional $J:E\to\mathbb{R}$ as

$$J(u)=\frac{1}{2}\sum_{t=1}^{T}|\Delta u(t-1)|^2-\frac{\lambda_k}{2}\sum_{t=1}^{T}|u(t)|^2-\sum_{t=1}^{T}F(t,u(t)),\quad u\in E, \tag{3.8}$$

where $F(t,x)=\int_0^x f(t,s)ds$, $f(t,x)=g(t,x)-\lambda_k x$. Then it is easy to show that J is of $C^2(E,\mathbb{R})$ with Fréchet derivatives given by

$$\langle J'(u),v\rangle=\sum_{t=1}^{T}[-\Delta^2u(t-1)+\lambda_k u(t)+f(t,u(t))]v(t). \tag{3.9}$$

Hence the solutions of (*P*) are exactly the critical points of *J* in *E*.

Lemma 3.1. [11] Let (1.1) and (f^+) (or (f^-)) hold, then the functional *J* defined by (3.8) satisfies (*C*) condition.

Lemma 3.2. [1] Let $h\in C([1,T]\times\mathbb{R},\mathbb{R})$ and $H(t,x)=\int_0^x h(t,s)ds$. Then the functional defined by

$$\breve{J}(u)=\frac{1}{2}\sum_{t=1}^{T}\left|\Delta u(t-1)\right|^{2}-\sum_{t=1}^{T}H(t,u(t))$$

satisfies (*PS*) condition in each of the following cases:

(i) h satisfies $h(t,x)=0$ for $x>0$ and $t\in[1,T]$ and

$$\lim_{x\to+\infty}\frac{h(t,x)}{x}=\alpha>\lambda_{1},\quad t\in[1,T];\tag{3.10}$$

(ii) h satisfies $h(t,x)=0$ for $x>0$ and $t\in[1,T]$ and

$$\lim_{x\to-\infty}\frac{h(t,x)}{x}=\alpha>\lambda_{1},\quad t\in[1,T].\tag{3.11}$$

Lemma 3.3. [11] Let (1.1) hold. We have

(i) $C_{q}(J,\infty)\cong\delta_{q,k-1}\mathbb{Z}$ provided (f^{-}) holds,

(ii) $C_{q}(J,\infty)\cong\delta_{q,k}\mathbb{Z}$ provided (f^{+}) holds.

Lemma 3.4. Let g satisfy (g^{+}) (or (g^{-})). Then J has a local linking at 0 with respect to the direct sum decomposition $E=E^{-}\oplus E^{+}$, where $E^{-}=span\{\phi_{1},\cdots,\phi_{m}\}$ (or $E^{-}=span\{\varphi_{1},\cdots,\varphi_{m-1}\}$ respectively), $E^{+}=(E^{-})^{\perp}$.

Proof. Suppose that (g^{+}) holds. Then

$$\frac{c_{1}}{p}\left|x\right|^{p}\leq G(t,x)-\frac{\lambda_{m}}{2}x^{2}\leq\frac{c_{2}}{p}\left|x\right|^{p},$$

where $p=\tau+1>2, G(t,x)=\int_{0}^{x}g(t,s)ds.$ Let $E^{-}=span\{\phi_{1},\cdots,\phi_{m}\}$. Then

$$E^{+}=span\{\phi_{m+1},\cdots,\phi_{T}\}.$$

For any $u\in E^{+}, \| u\|\leq\delta$ implies $|u(t)|\leq\delta$ for $t\in[1,T]$. Then we have

$$\begin{aligned}J(u)&\geq\frac{\lambda_{m+1}}{2}\sum_{t=1}^{T}|u(t)|^{2}-\frac{\lambda_{m}}{2}\sum_{t=1}^{T}|u(t)|^{2}-\frac{c_{2}}{p}\sum_{t=1}^{T}|u(t)|^{p}\\&=\frac{\lambda_{m+1}-\lambda_{m}}{2}\| u\|^{2}-\frac{c_{2}}{p}\| u\|_{p}^{p}\\&\geq\frac{\lambda_{m+1}-\lambda_{m}}{2}\| u\|^{2}-\frac{b_{p}c_{2}}{p}\| u\|^{p}.\end{aligned}$$

Noting that $p>2$, we can choose $0<\rho\leq\delta$ small enough such that $J(u)>0$ for $0<\| u\|\leq\rho, u\in E^{+}$.

For $u\in E^{-}$, similarly, $\| u\|\leq\delta$ implies $|u(t)|\leq\delta$ for $t\in[1,T]$. Thus we can get

$$J(u)\leq\frac{\lambda_{m}}{2}\sum_{t=1}^{T}|u(t)|^{2}-\frac{\lambda_{m}}{2}\sum_{t=1}^{T}|u(t)|^{2}-\frac{c_{1}}{p}\sum_{t=1}^{T}|u(t)|^{p}=-\frac{c_{1}}{p}\| u\|_{p}^{p}\leq 0.$$

This implies that J has a local linking at 0 with respect to $E = E^{-} \oplus E^{+}$. The case in which (g^{-}) holds can be similarly proved.

Now we give the proofs of Theorem 1.1--1.3.

Proof of Theorem 1.1. We only prove the case (i), the case (ii) can be similarly proved. By $g_x'(t,0) < \lambda_1$, a simple computation shows that u=0 is a local minimum of J. Hence

$$C_q(J,0) \cong \delta_{q,0}\mathbb{Z}, \qquad q \in \mathbb{N}. \tag{3.12}$$

Set

$$h_+(t,x) = \begin{cases} g(t,x), & x \geq 0, \\ 0, & x < 0, \end{cases}$$

and let $H_+(t,x) = \int_0^x h_+(t,s)ds$. Define the functional $J_+ : E \to \mathbb{R}$ as

$$J_+(u) = \frac{1}{2}\sum_{t=1}^{T} \left|\Delta u(t-1)\right|^2 - \sum_{t=1}^{T} H_+(t,u(t)), \quad u \in E, \tag{3.13}$$

Then the critical points of J_+ are exactly the solutions of the problem

$$(P_+)\begin{cases} -\Delta^2 u(t-1) = h_+(t,u(t)), & t \in [1,T], \\ u(0) = 0, \ \Delta u(T) = 0, \end{cases}$$

and the nonnegative solutions of (P_+) are the solutions of (P). By Lemma 3.2 (i), we see that $J_+ \in C^{2-0}(E,\mathbb{R})$ satisfies (PS) condition.

Since $g_x'(t,0) < \lambda_1, u = 0$ is a strictly local minimum of J_+. Hence there exist $\rho > 0, \tau > 0$ such that $J_+(u) \geq \tau, u \in E$ with $\|u\| = \rho.$ Since $\phi_1 > 0,$ by (1.1),

$$J_+(s\phi_1) \to -\infty, \quad s \to +\infty. \tag{3.14}$$

By Mountain Pass Theorem [14], J_+ has a nontrivial critical point u_+. Now we prove that $u_+(t) > 0$ for $t \in [1,T].$ Let $i \in [1,T]$ be such that $u(i) = \min_{j \in [1,T]} u(j).$ An immediate computation gives

$$\Delta u_+(i) \geq 0, \Delta u_+(i-1) \leq 0.$$

If $u_+(i) \leq 0,$ then $-\Delta^2 u_+(i-1) = 0.$ Hence $u_+(i-1) = u_+(i+1) = u_+(i) \leq 0.$ Repeating the process we can get

$$u_+(0) = u_+(1) = \cdots = u_+(T+1),$$

i.e., we get $u = 0.$ It contradicts that u_+ is a nontrivial critical point of J_+. Hence $u_+ > 0$ is a critical point of J. Moreover, by using the results in [7] and the critical group property for a mountain pass point [7], we have

$$C_q(J,u_+) \cong C_q(J_+,u_+) \cong \delta_{q,1}\mathbb{Z}, \quad q \in \mathbb{N}. \tag{3.15}$$

The same argument shows that *J* has a nontrivial critical point $u_- < 0$ with

$$C_q(J,u_-) \cong \delta_{q,1}\mathbb{Z}, \quad q \in \mathbb{N}. \tag{3.16}$$

By (f^+) and Lemma 3.3 (ii),

$$C_q(J,\infty) \cong \delta_{q,k}\mathbb{Z}, \quad q \in \mathbb{N}. \tag{3.17}$$

It follows from the relationship between the *q*-th Morse type number and the *q*-th Betti number that *J* must have a critical point u_0 such that

$$C_k(J,u_0) \neq 0. \tag{3.18}$$

Since

$$J''(u_0) = \begin{bmatrix} a_1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & a_2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & a_2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & a_{T-1} & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & a_T \end{bmatrix},$$

where

$$a_t = 2 - g_x'(t,u_0(t)), \quad t \in [1,T-1], \quad a_T = 1 - g_x'(T,u_0(T)), \quad \dim \ker(J''(u_0)) \leq 1,$$

by Shifting theorem and the critical groups characterizations of the local minimum and the local maximum [7],

$$C_q(J,u_0) \cong \delta_{q,k}\mathbb{Z}, \quad q \in \mathbb{N}. \tag{3.19}$$

It follows from $k \geq 2$ that u_+, u_- and u_0 are three nontrivial critical points of *J*. If $K = \{0, u_+, u_-, u_0\}$, then Morse equality (2.4) reduces to

$$(-1)^0 + (-1)^1 + (-1)^1 + (-1)^k = (-1)^k,$$

which is impossible. Thus, *J* must have the fourth critical point $u_1 \neq 0$. Therefore, u_+, u_-, u_0 are four nontrivial critical points of *J*.

Proof of Theorem 1.2. Since (g^+) (or (g^-)) implies $g_x'(t,0) = \lambda_m$ for all $t \in [1,T]$ or (1.2), $u = 0$ is a degenerate critical point of *J* with Morse index $\mu(0) = m-1$ and the nullity $\nu(0) = 1$. We only prove the case (i), the other cases can be proved in a similar way. By Lemma 3.4 and Proposition 2.1,

$$C_q(J,0) \cong \delta_{q,m}\mathbb{Z}, \quad q \in \mathbb{N}. \tag{3.10}$$

For $k = 1$, we get the following formulas by (3.17) and (3.18)

$$C_q(J,\infty)\cong\delta_{q,1}\mathbb{Z},\quad q\in\mathbb{N},\quad C_1(J,u_0)\neq 0.$$

Therefore u_0 is a mountain pass point of *J* and then

$$C_q(J,u_0)\cong\delta_{q,1}\mathbb{Z},\quad q\in\mathbb{N}. \tag{3.21}$$

As $m\neq 1, u_0\neq 0.$ If $K=\{0,u_0\}$, then the Morse equality (2.4) implies

$$(-1)^m+(-1)^1=(-1)^1.$$

It is impossible. Thus *J* has another nontrivial critical point u_1.

References

1. Agarwal, R.P., Perera, K., O'Regan, D.: Multiple positive solutions of singular discrete *p*-Laplacian problems via variational methods. Adv. Difference Equ. 2, 93–99 (2005)
2. Bartsch, T., Li, S.J.: Critical point theory for asympotically quadratic functional and application to the problems with resonance. Nonlinear Anal. 28, 419–441 (1997)
3. Bonanno, G., Candito, P.: Infinitely many solutions for a class of discrete non-linear boundary value problems. Appl. Anal. 88, 605–616 (2009)
4. Bonanno, G., Candito, P.: Nonlinear difference equations investigated via critical point methods. Nonlinear Anal. 70, 3180–3186 (2009)
5. Candito, P., Giovannelli, N.: Multiple solutions for a discrete boundary value problem involving the *p*-Laplacian. Comput. Math. Appl. 56, 959–964 (2008)
6. Cerami, G.: An existence criterion for the critical points on unbounded manifolds. Isit. Lombardo Accaqd. Sci. Lett. Rend. A 112, 332–336 (1978) (in Italian)
7. Chang, K.C.: Infinite Dimensional Morse Theory and Multiple Solutions Problems. Birkhauser, Boston (1993)
8. Chang, K.C.: H^1 versus C^1 isolated critical points. C. R. Acad. Sci. Paris 319, 441–446 (1994)
9. Jiang, L., Zhou, Z.: Existence of nontrivial solutions for discrete nonlinear two point boundary value problems. Appl. Math. Comput. 180, 318–329 (2006)
10. Li, S., Liu, J.Q.: Some existence theorems on multiple critical points and their applications. Kexue Tongbao 17, 1025–1027 (1984)
11. Li, S.J., Willem, M.: Applications of local linking to critical point theory. J. Math. Anal. Appl. 189, 6–32 (1995)
12. Liu, J.Q.: A Morse index for a saddle point. Syst. Sci. Math. Sci. 2, 32–39 (1989)
13. Liu, J.S., Wang, S.L., Zhang, J.M.: Multiple solutions for boundary value problems of second-order difference equations with resonance. J. Math. Anal. Appl. 374, 187–196 (2011)
14. Rabinowitz, P.: Minimax methods in critical point theory with applications for differential equations. In: CBMS Regional Conference (1984)
15. Wang, D., Guan, W.: Three positive solutions of boundary value problems for *p*-Laplacian difference equations. Comput. Math. Appl. 55, 1943–1949 (2008)
16. Zhu, B.S., Yu, J.S.: Multiple positive solutions for resonant difference equations. Math. Comput. Modelling 49, 1928–1936 (2009)

The Application of the GPRS Network on the Design of Real-Time Monitor System for Water Pollution Resource

Shi-he Sun

Laiwu Vocational and Technical College, Laiwu, Shandong, China
Sshh0518@sina.com

Abstract. In order to design the real-time monitor system with good performance, the application of GPRS network on it was studied in depth. Firstly, the importance of designing a real-time monitor system was introduced, and the basic theory of GPRS network was explained; and then the theory principle of real-time monitor system of water pollution resource was analyzed. And then methodology of GPRS network was studied. Finally the hardware device, software and theirs corresponding function were designed. And the application of this monitor system of water pollution resource was good when it was applied in actual engineering.

Keywords: GPRS, real-time, Monitor system, water pollution resources.

1 Introduction

Monitor of water pollution resources could grasp and understand the condition and tendency of waste discharge, and the monitor results and materials were main basis for executing environmental laws and standard, and develop environmental management.

Real-time monitor system of water pollution resources based on GPRS was a comprehensive online instrument which was made up of analysis software and communication network used modern sensor technology, automatic measuring technology, automatic controlling technology, computer technology and other relating technologies. Real-time automatic monitor system of water pollution resources was made up of field monitor station system, data transfer system, monitor center of water pollution resources and remote monitor and management system of water pollution resources system. Monitor system could monitor data of water quality constantly and dynamically, and the statistical operations and deals with data were carried out, and formed all kinds of reports and figures and tables, and the monitor results were transferred to monitor center and environment protect policy.

The full name of GPRS was General Packet Radio Service, which was a system made up of GSMPhase2 and standard realization, and it could offer quick data transfer efficient. GPRS network could offer end-to-end, big field wireless IP joints, and it had excellent goodness in a lot of engineering application fields. The dynamic management system of water pollution resources applied a lot of virtues of GPRS network, for example, more quick speed and quick joint, in condition that the need put forward and was controlled by a good many generating data. GPRS network could offer real-time wireless transfer and was very fast without a dia1-up modem connection to GPS

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 65–71, 2011.

machine for resource data. These content was very critical mainly because there was a little data room in GPS position information, and constant transmission should be needed. Therefore the system of monitor system for water pollution resources could apply GPRS network to transfer GPS position information and relating information.

GPRS network was an advanced technology for the Global System Mobile Communication (GSM), which would amend and simplify the wireless relating to packet data networks and so on. GPRS adds two nodes to GSM, which were SGSN node and gateway GPRS supporting node and used a packet radio rule to efficiently transmission client data packets between mobile stations and externa1 networks including packet data.

GPRS would make radio resources used efficiently, and got packet data easily, and provide better transmission speed and billing based on volume. Recently the GPRS which was established by European Telecommunication Standards Institute attracted many scientists in these study fields.

Recently, monitor methods of water pollution resources had two kinds, which were telephone line transfer and manual copy down; these methods had bad real time capability, and would cost most money, however GPRS network had the following advantages:

(1) GPRS users could casually distribute and mobile theirs network point, and could not worry about the maintenance of lines, or communications were brought in radio contact. When the new monitor point was established the lines needn't to be distributed and other operations. Less investment was used in GPRS network than optical fiber device and leased line system. And it was installed fast.

(2) Price of terminal unit was relative low. Price of its terminal unit was cheaper than that of DTU, DDN leaser line Modem.

(3) Charges of GPRS network were cheaper, and accounting cost of GPRS networks was reasonable. Monthly payment of GPRS network was cheaper than satellite telephone network. In GPRS network, users only established connections with network once, and would keep these connections for a long time, and the accounting would began when the data was transferred and the communication channel was occupied. Therefore the monitor point need not be connected continually; and the charges of transfer gap could not be accounted.

(4) GPRS could support the continually, light abrupt data service. The communication quality was stable and reliable, and could not fall line.

(5) GPRS could connect with network quickly, and offer seamless connection present data network. Because GPRS network was a group digital network and it support TCP/IP, X.25 protocol, therefore it could be connected with group data network directly and need not be switch connected with PSTN and other networks, the connection speed was only a several second, and this speed was more fast than circuit data service.

TCP was intended for use as a highly reliable host-to-host protocol between hosts in packet-switched computer communication networks, and in interconnected systems of such network. GPRS main element was listed in Table 1.

Table 1. GPRS main element

Node	Agent	Queue
MS	MSAgent	Um_DL_Queue
BSS	BSSAgent	Gb_DL_Queue
GGSN	GGSNAgent	Gi_DL_Queue

2 Rule of Real Time Monitor System of Water Pollution Resources

(1) Structure of real-time monitor system of water pollution resources

Real-time monitor system of water pollution resources was made up of mobile terminals derived by equipment, a dispatch center and GPRS network, which was showed in Figure 1.

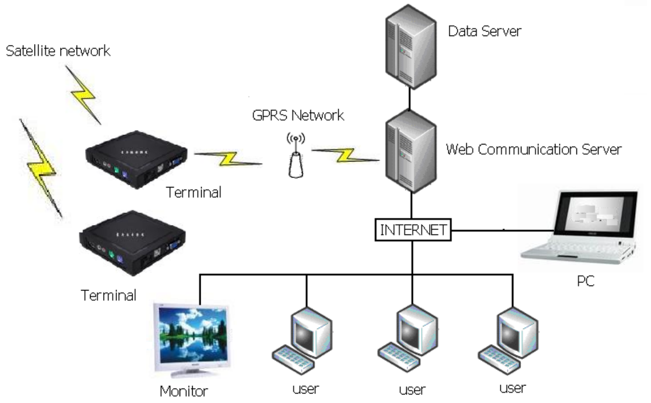


Fig. 1. The structure figure of GPRS network for monitoring water pollution resources

The mobile terminals could got GPS information and compute the data relating to water pollution resources. The signals of water pollution resources could be got through connecting the expansion interfaces of mobile terminals in the monitor system to a lot of control lines and examination. All kinds of data could be imported to the monitor center using GPRS and internet.

GPRS was a communication network could transfer the data of water pollution resources, which was in the middle of dispatch center and mobile terminals, such as alarm data of water pollution resources to the dispatch center, and then the control commands could be imported to the monitor system.

3 Linear Programming Model of Real Time Monitor System

(1) Modeling of real-time monitor system

A model of monitoring water pollution resources was established for real time monitor system. The number of water pollution resources was defined as m, the number of monitor station of water pollution resources was defined as n, the number of water pollution resources (i) to the number of monitor station (j) was defined as x_{ij}, the water quality level (f) could be got according to water pollution resources (i) was defined as g_{fi}, the monitor station quality level (s) could be got through the monitor station (j) was defined as g_{si}, the corresponding constraints were defined as following:

(a) The number of water pollution resources and water quality level. The number of water pollution resources monitored by monitor station should be within the scale of capability of monitor station, and the following conditions should be meeting:

$$Q_i \le \sum_{i=1}^{m} x_{ij} \le A_i,\ (j = 1,2,\cdots m) \tag{1}$$

where x_{ij} was the number of water pollution resources (i) for monitor stations (j), A_i was maximum number of water pollution resources, and Q_i was the capability of monitor stations.

(2) Minimum task of monitor station of water pollution resources. In the environmental project, every monitor station could be given a minimum monitoring work. Assume that the minimum monitor work (r) of monitor station (j) was Q_{jr}, the following equation was could be got:

$$\sum_{i=1}^{m} x_{ij} \ge Q_{jr} \tag{2}$$

(3) Non-negative constraint conditions. The number of water pollution monitor offered by a monitor station could not be negative; the following condition could be needed:

$$x_{ij}\ (i = 1,2,\cdots,m;\ j = 1,2,\cdots,n) \tag{3}$$

(4) Objection equation. The target water pollution level of the water pollution resources was defined as g and its mistake scale should be not more than 5%. Based on the actual status of water pollution resource, the error from real water pollution resources and object water pollution level at every monitor station of water pollution resources should be minimized by the objective function, ie.

$$\min\ s = \sum_{j=1}^{n} \left| (Q_1 \cdots Q_{j-1} Q_{j+1} \cdots Q_n) \sum_{i=1}^{m} x_{ij} g_{fi} \right| \tag{4}$$

where, g_{fi} denoted the water pollution resources level offered by water pollution resources (i) and Q_j the task of monitor station (j). Based on the real water pollution resources requirement, the above mathematics model was improved through giving a new constrain.

$$\left|\sum_{i=1}^{m} x_{ij} g_{fi}\right| \geq Q_i g_{sj} \tag{5}$$

(2) The solution of mathematical model

The above model was coped with two step means, In the first step, there were many new more than or equal to zero variables, which were defined as (x_{n+1} , x_{n+2} ,…, x_{n+h}), were amended in the mathematical model. The object of this operation was to make the m-moment unit sub matrix concluded coefficient matrix A , which could be expressed as follows:

$$A = (b_{ij})_{m\times(n+h)} \; i = 1,2,\cdots,n \; ; \; j = 1,2,\cdots,n+h \tag{6}$$

In the first step, the summary of every amended artificial variable was decreased, the corresponding function could be expressed as following:

$$\min \; Z_1 = \sum_{i=n+1}^{n+h} x_i \tag{7}$$

If the good resolution of Z_1 =0 could be got. Every amended artificial variable were non-basic variables, but the m original variables, before the addition, were basic variables. This left an m-moment unit sub matrix in the coefficient matrix when the corresponding column of artificial variables was deleted and it was assumed that this was the initial feasible base B_0 . Then we turned to the second step of the two-stage means to deal with the mathematics equation. Or there was no optimal resolution for dealing with this mathematics problem.

4 Methodology of GPRS Network

The standard TCP receiver formed reset packets when packets were received through an aborted joint. An optimal algorithm was put forward for keeping back the delivery of aborted information over the last-hop link Um_DL_Queue, which was defined as Fast Reset. When SGSNAgent got a reset packet in up link (UL), decreased every packet that was the flow to that of the reset packet in SGSNAgent and BSSAgent. Therefore, these unnecessary packets need not be transferred to MS keeping up limited radio bandwidth and battery power of M S.

If a DL packet's fields were same as that in the stored information database of its session but for the sequence account was less than the acknowledgement amount, therefore this duplicate packet could decreased.

UL IP and TCP packet-fields stored by SGSNAgent and DL IP and TCP packet-fields to be contracted were listed in Table 2 and Table 3.

Table 2. UL IP and TCP packet-fields stored by SGSNAgent

Field	Length
IP destination address	64bit(ipv4)/512bit(ipv6)
IP source address	64bit(ipv4)/512bit(ipv6)
TCP destination port	32bit
TCP source port	32bit

Table 3. DL IP and TCP packet-fields to be contracted

Field	Length
IP destination address	64bit(ipv4)/512bit(ipv6)
IP source address	64bit(ipv4)/512bit(ipv6)
TCP destination port	32bit
TCP source port	32bit

5 Monitor Centers of Water Pollution Resources

Monitor applied standard C/S system structure, and the general software and hardware production were used, and the stored formula was ruled, and the compatibility of monitor system was higher, and scale of system was easy to be extended. The basic function was listed as follows:

(1) Center controlling function: in the monitor center the real time data could be got through the real time monitor of network;

(2) Center alarm function: the abnormal alarm signals in field could be got through sound and light in the monitor center;

(3) Data stored function: the original data, alarm data and operation information could be stored according to the database formula.

(4) Data distribution function: center database could achieve the share of data, and had opening characteristic.

6 Hardware Device and Corresponding Function

(1) Server: server was a core device of the center monitor, SCADA control unit and relationship database was run. According to different functions the data server, history server and WEB database server could be assigned real-time. And the main server used redundant allocation, which ensured that any fault of a server could not make the whole system run failure.

(2) Engineer and monitor station: monitor operation station was a joint between man and machine of system, it was used as user machine in the center monitor station, and it could offer intuitional monitor figure and data show, operators could grasp the monitor status of all the system and put a command out, which was connected with server through LAN and exchanged information.

(3) Network device: the main assignment had Ethernet exchanger, router, firewall, and so on.

7 Application Software of Monitor System

(1) Communication software of GPRS data collecting device
(2) Internet center server data collecting and distributing software, data processing software and Web server software.
(3) Working station monitor software and long distance data maintains software.

Communication software of GPRS data collecting device was installed in the PLC before the data collecting device and terminals in the GPRS wireless network, and it could achieve the installment and removing, compression, encryption and transmission of the joint protocol and so on. The main function of Internet center server data collecting and sending software, data processing software and Web server software was listed as follows: based on SOCKET communication joint, TCP/IP protocol, the center server used as communication bridge, collecting data and running windows high level service SQL SERVER, which would process data, and at the same time transferred controlling commands, and ensured the normal communication between the field and user. Web Server program running IIS application, and make the legal users find data using IE browser at any time and any location. Client application program could achieve the following functions: the SOCKET joint and SQL SERVER joint were constructed with center server. And at the same time the online and real-time running status of all drinking water sources was showed, and the running parameters, history data and report table input was checked. When alarm condition was set and the system could send alarm status to the email box and cell phone. Clients could control the drinking water sources if the data was sufficient, and could carry out long distance maintenance for center server.

8 Conclusions

The monitor system of water pollution resources had been applied in the actual application, and the data transfer clearly, safely and correctly; the devices of system run reliably; and the fault ratio was low, the maintenance was proper, and the monitor projects were complete, and the data had higher credibility; and the automatic monitor function of sewage was carried out sufficiently, and the data could be provided to many related department, and the water pollution could be prevented effectively.

References

1. Qiu, Q.-l., Zhang, D.-m., Ma, J., et al.: Redundancy Elimination in GPRS network. Journal of Zhejiang University SCIENCE A 17(4), 447–482 (2006)
2. Gu, Q.-h., Lu, C.-w., Guo, J.-p., et al.: Dynamic management system of ore blending in an open pit mine based on GIS/GPS/GPRS. Information & Management 20, 0132–0137 (2010)
3. Wang, J., Gao, J.-X., Wang, J.-L., Xu, C.-H.: EMD-based GPS baseline solution and validation test. Journal of Cina University of Mining and Technology 118(2), 283–287 (2008)
4. Gu, Q.-H., Lu, C.-W., Li, F.-B., Wan, C.-Y.: Monitoring dispatch inform ation system of trucks and shovels in an open pit based on GIS/GPS/GPRS. Journal of Cina University of Mining and Technology 18(2), 288–292 (2008)

CuttingPlane: An Efficient Algorithm for Three-Dimensional Spatial Skyline Queries Based on Dynamic Scan Theory

Meng Zhao and Jing Yu

Department of Computer Science and Technology
Yanshan University
Qinhuangdao 066004, China
Zhaomeng_527@163.com, xyyj@ysu.edu.cn

Abstract. Skyline operator and skyline computation play an important role in database communication, decision support, data visualization, spatial database and so on. In this paper we firstly analyze the existing methods, point out some problems in progressive disposal, query efficiency and convenience of following user selection. Secondly, we propose and prove a theorem for pruning query space based on dynamic scan theorem, based on the thought of the theorem, we propose a more efficient algorithm-dynamic cutting plane scan queries for skyline queries, and analyze and verify the feasibility, efficiency and veracity of the algorithm through instance and experiment.

Keywords: skyline, dominate, dynamic scan.

1 Introduction

The integration of position locators and mobile devices enables new pervasive location-aware computing environments where all objects of interest can determine their locations. In such environments, moving objects move continuously and send location updates periodically to spatial databases. Spatial database servers index the locations of moving objects and process outstanding continuous queries. Characterized by a large number of moving objects and a large number of continuous spatial queries, spatiotemporal databases are required to exhibit high scalability in terms of the number of moving objects and the number of continuous queries.

To increase the scalability of spatial databases, there exist two main challenges. The first challenge is to support a large set of continuous queries concurrently. With the ubiquity and pervasiveness of location-aware devices and services, a set of continuous queries execute simultaneously in a spatial database server. In the case that the number of queries is too large, the performance of the database degrades and queries suffer long response time.

Because of the real-timeliness of the location-aware applications, long delay makes the query answers obsolete. Therefore, new query processing algorithms addressing both efficiency and scalability are required for answering a set of concurrent spatial queries.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 72–78, 2011.

The second challenge for building scalable spatial databases is to index moving objects efficiently. Building indexes on moving objects can facilitate significantly query processing in spatial databases. However, due to the dynamic property of moving objects, the underlying indexing structures will receive numerous updates during a short period of time. Given the fact that update processing is costly, traditional spatial indexes may not be applied directly to spatial databases. This situation calls for new indexing techniques supporting frequent updates.

The above two challenges motivate us to develop scalable techniques for both continuous query processing and moving object indexing in spatial databases.

2 Related Work

Skylines, and some directly related problems such as multi-objective optimization[3], maximum vectors[4,5,6] and the contour problem[7], have been extensively studied and numerous algorithms have been proposed for main memory processing.

Borzsonyi et al. propose block nested loop approach(BNL)[8]. A straightforward approach to compute the skyline is to compare each point p with every other point; if p is not dominated, then it is a part of the skyline. BNL builds on this concept by scanning the data file and keeping a list of candidate skyline points in main memory. The advantage of BNL is its wide applicability, since it can be used for any dimensionality without indexing or sorting the data file. Its main problems are the reliance on main memory and its inadequacy for on-line processing.

Divide-and-Conquer approach[8](D&C) divides the dataset into several partitions so that each partition fits in memory. Then, the partial skyline of the points in every partition is computed using a main-memory algorithm[5,6] and the final skyline is obtained by merging the partial ones. D&C is efficient only for small datasets. For large datasets, the partitioning process requires reading and writing the entire dataset at least once, thus incurring significant IO cost.

Bitmap approach[9] encodes in bitmaps all the information required to decide whether a point is in the skyline. The efficiency of bitmap relies on the speed of bit-wise operations. The approach can quickly return the first few skyline points according to their insertion order (e.g., alphabetical order), but cannot adapt to different user preferences, which is an important property of a good skyline algorithm[10]. Furthermore, the computation of the entire skyline is expensive because, for each point inspected, it must retrieve the bitmaps of all points in order to obtain the juxtapositions.

Kossmann et al. present NN(Nearest Neighbor) approach[10] due to its reliance on nearest neighbor search, which applies the divide-and-conquer framework on datasets indexed by R-trees[11]. NN uses the results of nearest neighbor search to partition the data universe recursively. NN performs a nearest neighbor query using an existing algorithm[12] on the R-tree, to find the point with the minimum distance (mindist) from the beginning of the axes (point o). NN has also some serious shortcomings such as need for duplicate elimination, multiple node visits and large space requirements.

3 Dynamic Scanning Cutting Plane Algorithm

3.1 Formal Problem Definition

Assume that we have a database of N objects. Each database object p with d real-valued attributes can be conceptualized as a d-dimensional point (p1, . . . , pd) □ Rd where pi is the i-th attribute of p. We use P to refer to the set of all these points. As shown in the coordinate system of figure 1, The coordinates of the system represent the description of a hotel with three attributes: distance to beach, the quality and price of the hotel. the coordinates of point S is less than any point in the coordinate system area D it constructed, which parallels with the initial coordinate system constructed by point O. The area D is called Dominance region of point S. As shown in figure 2, points a, b, c are not dominated by any point in coordinate system constructed by point O. Therefore points a, b, c construct the skyline, expressed as skyline={a, b, c}. The points a, b, c are called skyline points.

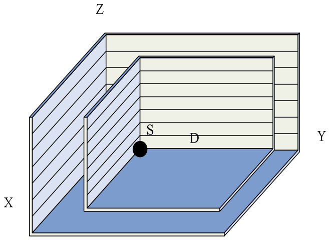


Fig. 1. Dominance region

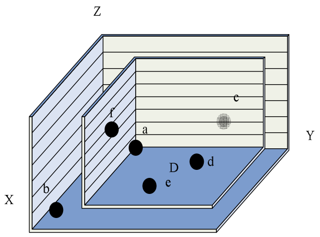


Fig. 2. Example of skyline

According to the example of figure 2, there are many hotels around traveler, who is traveling in a foreign country. Some hotels may be closer to the traveler, but the price of room may be more expensive, and there may be no room. How to select the best scenario for traveler? The skyline computation can solve this easier. As shown in figure 2, make the traveler statement as origin point, and construct the coordinate. X axis of coordinates denotes the distance between hotel and the traveler; Y axis of coordinates denotes the price of the hotel; Z axis of coordinates denotes the number of people who get accommodation in the hotel then, in other words, more people get accommodation, less rooms the hotel left. Points a, b, c are skyline points, Point d, e, f are dominated by point a. By skyline computation, return skyline points a, b, c to traveler. Then traveler considers his condition and makes decision.

According to analyzing current skyline query methods, there are some shortages, such as progressive disposal of skyline points, computation of the points in spatial database, the minimal set of the best candidates for the skyline query and so on, a new idea for three-dimensional points in spatial database was proposed, the core of the idea is that a dynamic scan cutting plane is initialized, then moves towards the direction of the vector {1, 1, 1}, and scans spatial data points, according to the definition of skyline region domination, filtrate useless spatial points and return the skyline points. The method can progressively feedback skyline points by ascending order of any dimension.

There are several advantages for three-dimensional spatial skyline queries based on dynamic scan theory: filtrates as many points as it can, saves as much costs as it can; progressively feedbacks skyline points by ascending order, convenient for the user decision; some skyline points can be got when the user terminates the query, these points that the dynamic scan cutting plane had scanned are part of the whole skyline.

Based on the dynamic scan cutting plane theory, the algorithm called CuttingPlane is proposed.

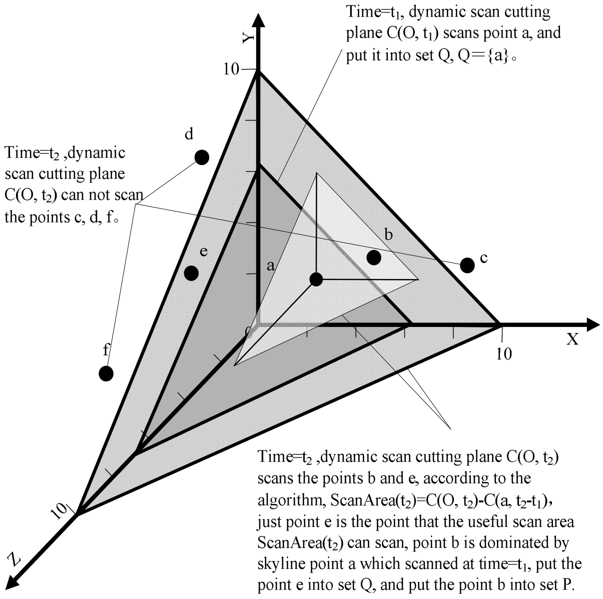
3.2 Algorithm Description

The algorithm CuttingPlane is described concretely as follows:

```
Input: the set of spatial data points S, the speed of dynamic scan cutting plane v
Output: the set of skyline points Q
Algorithm CuttingPlane (S, v)
{Q=NULL;
//The set of skyline points Q is initialized by Null P=NULL;
//The set of medial filtered points is initialized by Null
FrameofAxes(S);
//According to the input set S, use function FrameofAxes to construct three-dimensional coordinate system, initialize the spatial data points in coordinate system, and the origin point O is returned.
C(O, 0)=construct(O, 0, 60, v);
//Time=0, the dynamic scan cutting plane is initialized by function Construct(Origin, time, angle, speed), which is 60° angle with the coordinate system, C(O, 0) is the area that cutting plane cuts the coordinate system constructed by the origin point O, the dynamic scan cutting plane moves by the speed v and towards the direction of the vector {1, 1, 1}.
ScanArea(0)=C(O, 0);
//The initiative useful area is initialized by C(O, 0)
// The process of the dynamic scanning
While(S!=Q ∪ P)//The scanning will terminate when Q ∪ P=S.
{ If(Q!=NULL) Printf(Q); //Print the skyline points in the process of the dynamic scanning
Move(C(O, t), v);
//the dynamic scan cutting plane moves by the speed v and towards the direction of the vector {1, 1, 1}.
If(when time=t, dynamic scan cutting plane C(O, t) scans points)
{If(Q!=NULL)
{For(i=0; i<=Num(Q); i++)
{ScanArea(t)=C(O, t)-C(Q[i], t-Scantime(Q[i]));}}
//When time=t, the useful scan area ScanArea(t) of C(O, t) is the left area that C(O, t) wipes off the areas, which are dominated by the points of the skyline points set Q.
If(when time=t, the useful scan area ScanArea(t) of C(O, t) scans points)
{M=GetPoints(ScanArea(t));
//Function GetPoints() gets the points scanned by the useful scan area ScanArea(t)
Q=Q ∪ M; //Put the points scanned by the useful scan area ScanArea(t) into set Q
For(i=1; i<=num(M); i++)
P=P∪Contain(M[i]);
//Use the function Contain() gets the points dominated by the set of M, and puts
them into set P
  } }
return(Q);
}
```

3.3 Example Analysis

As shown in figure 3, the example explains the query progress of the algorithm CuttingPlane.



Time=t_1, dynamic scan cutting plane C(O, t_1) scans point a, and put it into set Q, Q={a}。

Time=t_2 ,dynamic scan cutting plane C(O, t_2) can not scan the points c, d, f。

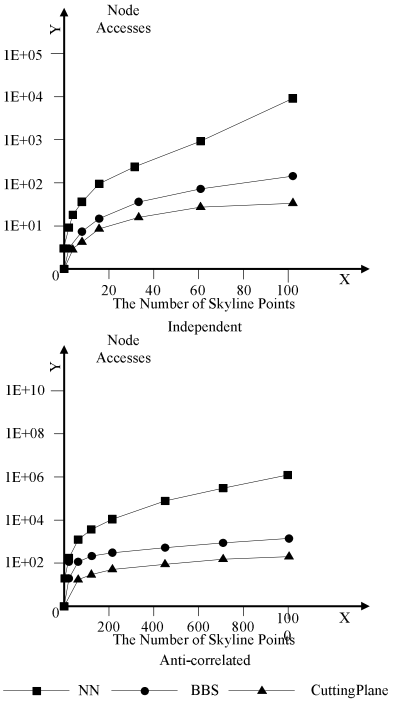
Time=t_2 ,dynamic scan cutting plane C(O, t_2) scans the points b and e, according to the algorithm, ScanArea(t_2)=C(O, t_2)-C(a, t_2-t_1), just point e is the point that the useful scan area ScanArea(t_2) can scan, point b is dominated by skyline point a which scanned at time=t_1, put the point e into set Q, and put the point b into set P.

Fig. 3. The process of progressive returning skyline points

Input the set of spatial data points S={a, b, c, d, e, f}, the speed of dynamic scan cutting plane v, firstly use function FrameofAxes(S) to construct the coordinate system, and return origin point O; Secondly time=0, the dynamic scan cutting plane is initialized by function Construct(Origin, time, angle, speed), which is 60° angle with the coordinate system, C(O, 0) is the area that cutting plane cuts the coordinate system constructed by the origin point O, the dynamic scan cutting plane moves by the speed v and towards the direction of the vector {1, 1, 1}, and initialize the useful area of the cutting plane, ScanArea(0)=C(O, 0); Thirdly do dynamic scan progress, time=t1, dynamic scan cutting plane scans point a, because the set of skyline points Q is NULL at the time, the useful area of the cutting plane ScanArea(t1)=C(O, t1), put the point a into set Q, Q={a}, use function Contain() to find the points that point a dominates, and put them into set P, P={b, c}; The scanning moves on, time=t2, ScanArea(t2)=C(O, t2)-C(a, t2-t1), ScanArea(t2) scans the point e, put it into the set Q, Q={a, e}, use function Contain() to find the points (d, f) that point e dominates, and put them into set P, P={b, c} ∪ {d, f}={b, c, d, f}, and at the time, Q ∪ P={a, b, c, d, e, f}=S, the dynamic scan progress is terminated, the set of the skyline points Q is returned.

4 Experimental Analysis

All our experiments are carried out on a Sun Ultra Workstation with a 2.4GHz processor and 1 GB of main memory. The benchmark databases and intermediate query results are stored on a 120GB Seagate disk drive. For our experiments, we implemented the NN, BBS and CuttingPlane algorithms in VC. In order to study the effect of dimensionality we use the datasets with cardinality N=1M and vary d=3. Figure 4 and 5 show the number of node accesses and CPU time.



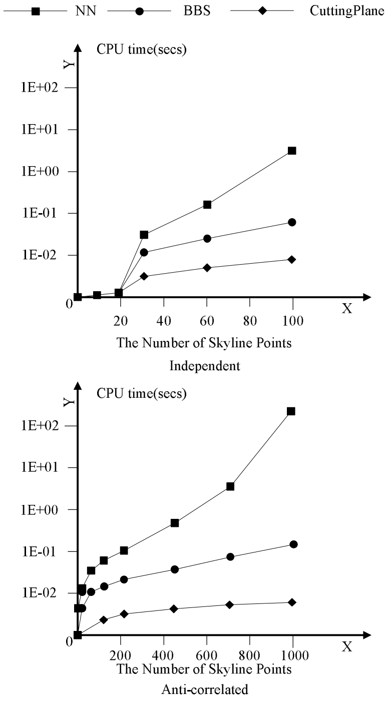


Fig. 4. The comparison of node accesses **Fig. 5.** The comparison of CPU time

From the experiments, we can clearly see the effectiveness and efficiency of CuttingPlane by comparing it against NN and BBS.

5 Conclusion

This algorithm CuttingPlane based on dynamic scan retrieves the objects of spatial dataset adopting the skyline query, and then uses the geometry principle and the skyline dominate region theory to get the data points satisfying the skyline conditions. By moving the dynamic scan cutting plane continuously we can prune the search space, it is not necessary to retrieve all the points, so it can reduce the number of the search points and decrease the query cost. We apply dynamic scan cutting plane query on datasets indexed by R-trees, and get the relation of the objects and search area. The improvement of algorithm for skyline queries will promote the development of the database community continuously.

References

1. Tan, K., Eng, P., Ooi, B.: Efficient Progressive Skyline Computation. In: VLDB (2001)
2. Borzsonyi, S., Kossmann, D., Stocker, K.: The skyline operator. In: Proc. IEEE Conf. on Data Engineering, Heidelberg, Germany (2001)
3. Steuer, R.: Multiple Criteria Optimization. Wiley, New York (1986)
4. Kung, H., Luccio, F., Preparata, F.: On Finding the Maxima of a Set of Vectors. Journal of the ACM 22(4) (1975)
5. Stojmenovic, I., Miyakawa, M.: An Optimal Parallel Algorithm for Solving the Maximal Elements Problem in the Plane. Parallel Computing 7(2) (1988)
6. Matousek, J.: Computing Dominances in En. Information Processing Letters 38(5) (1991)
7. McLain, D.: Drawing Contours from Arbitrary Data Points. Computer Journal 17(4) (1974)
8. Borzsonyi, S., Kossmann, D., Stocker, K.: The Skyline Operator. In: ICDE (2001)
9. Tan, K., Eng, P., Ooi, B.: Efficient Progressive Skyline Computation. In: VLDB (2001)
10. Kossmann, D., Ramsak, F., Rost, S.: Shooting Stars in the Sky: an Online Algorithm for Skyline Queries. In: VLDB (2002)
11. Guttman, A.: R-Tree a dynamic index structure for special search. In: Proc. ACM SIGMOD, pp. 47-57 (1984)
12. Roussopoulos, N., Kelly, S., Vincent, F.: Nearest Neighbor Queries. In: SIGMOD (1995)

ROS: Run-Time Optimization of SPARQL Queries

Liuqing Li, Xin Wang*, Xiansen Meng, and Zhiyong Feng

School of Computer Science and Technology, Tianjin University, Tianjin 300072, China
fs19861226@163.com, {wangx,zyfeng,laomeng_boy}@tju.edu.cn

Abstract. The optimization effect on large-scale RDF data is not satisfactory using the existing algorithms based on cost models. This paper presents the Run-time Optimization of SPARQL queries (ROS), and describes the join graphs and the index structures for SPARQL queries that are foundations of the ROS approach. The ROS algorithm, without cost models, intertwines cost estimation and query optimization into the execution procedure, and determines query plans in run time. Our experiments using the SP2Bench benchmark show that ROS can select the best query plan and improve query efficiency dramatically compared with the existing approaches.

Keywords: SPARQL, query optimization, run time, join graph, RDF.

1 Introduction

As increasingly large RDF datasets are being published on the Web, effcient RDF querying has become an essential factor in realizing the Semantic Web vision [1]. To facilitate RDF data access, the W3C has standardized the SPARQL query language, which is based upon a powerful graph pattern matching facility. In order to improve query efficiency, a query optimization algorithm is needed. The target of optimization algorithms are to determine the optimal query plan for execution from all candidate query plans.

The existing query optimization algorithms can be classified into three categories: (1) The parameter-based approaches [2][3], without statistical information, which estimate query plan execution cost by evaluating a function that has parameters. These approaches may make large errors if no law can be detected on the distribution of data. (2) The sample-based approaches [4][5], also without statistical information, which extract representative sample data directly from the dataset, select the least cost solution by comparing the costs of query plans that using the sample data as the input. The disadvantage of these approaches is that it is difficult to get the most representative sample data that is used to ensure the accuracy of query plan costs. (3) The statistical-based approaches [6], which evaluate costs using statistical information. The accuracy of these approaches depends on the accuracy of statistics, but it will take a lot of time to calculate all necessary statistics.

* Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 79–86, 2011.

Dynamic query optimization approaches in [7] select a plan from a number of options in the compilation phase based on cost models. After execution for a period of time, they will do re-selection according to the intermediate results. These approaches need accurate cost models. The recently proposed ROX in [8] can overcome the disadvantages of conventional query optimization approaches. Without accurate cost models, ROX simplifies the compilation phase, puts selection of query plans into run time, and intertwines optimization and estimation into query execution.

In this paper, we propose a novel approach to Run-time Optimization of SPARQL queries (ROS). The ROS approach limits static compilation to normalization, simplification and the identification of join graphs. The join graphs cluster SPARQL join operators. ROS executes operations in the join graph one by one, fully materializing intermediate results. It is crucial to recognize that ROS is not merely running a query optimizer at run-time, but also intertwines and integrates query optimization into query evaluation. Our experimental results show that ROS can consistently find the good plan from the search space.

The contributions of this paper can be summarized as follows: (1) ROS does not need a precise cost model. It does not have to spend extra time in building a cost model, which can not be accurate actually. (2) The resulting ROS optimizer obviously improves the query efficiency.

In Section 2, we introduce the basic knowledge of ROS, which includes the definition of the join graph and the description of the index structure. Section 3 describes the ROS algorithm in detail. In Section 4, we extensively test ROS, showing that it can find better query plans than the approach adopted by Sesame. Finally, in Section 5, we conclude this paper and outlook the future work.

2 Preliminaries

In this section, we will introduce formal definitions and symbols of the ROS algorithm. The SPARQL query language includes many operators, but the most important one is the join operator. ROS is aimed at optimizing the join order in run time so as to improve the efficiency of the join operator. First, a sequence-independent representation of a query plan in run time is shown, which is a similar structure like the join graph defined in [8]. Then, we will introduce the index structure used by ROS.

2.1 Join Graphs

We first give the related definitions of a join graph and then describe with more details its vertex and edge components.

Defination 1. Let $var = (value, name)$, where *name* means a local variable name and *value* its value, *var* is divided into two categories: (1) if *var* is a constant, then its value is self-increasing. (2) if *var* is a variable, then it begins with the ? symbol.

Defination 2. Let *V* be a finite set of *var*s and $T = (subject, predicate, object)$ be a triple pattern, where *subject*, *predicate*, $object \in V$.

For instance, in a triple pattern (?article, rdf:type, bench:Article), ?article represents a variable and is also the subject, rdf:type represents a constant and is also the predicate, bench:Article represents a constant and is also the object.

Defination 3. A *join graph* $G=(V,E)$ is defined as an edge-labled graph where: (1) A vertex $v \in V$ represents a SPARQL triple pattern or a SPARQL join operator. (2) An edge $e \in E$ represents a join operation or two triple vertices that have common variables.

Join graphs are obtained through the following steps: First, we get a SPARQL syntax tree through the compilation phase. The tree is an initial plan after static optimization in the compilation phase. Then, we rewrite the plan and move out the *distinct* and *sort* operators through the equivalent transformation. We make up a plan containing only *projection*, *join* and other operators. Fig.1 shows two SPARQL queries Q1 and Q2.

```
SELECT DISTINCT ?name
WHERE {
   ?article   rdf:type    bench:Article .
   ?article   dc:creator   ?person .
   ?inproc    rdf:type    bench:Inproceedings .
   ?inproc    dc:creator   ?person .
   ?person    foaf:name    ?name
}
```

(a) Q1

```
SELECT DISTINCT ?name1 ?name2
WHERE {
   ?article1   rdf:type    bench:Article .
   ?article2   rdf:type    bench:Article .
   ?articel1   dc:creator ?author1 .
   ?author1    foaf:name ?name1 .
   ?articel2   dc:creator ?author2 .
   ?author2    foaf:name ?name2 .
   ?articel1   swrc:journal ?journal .
   ?articel2   swrc:journal ?journal
}
```

(b) Q2

Fig. 1. SPARQL queries Q1 and Q2

Fig.2 is the join graph of SPARQL query Q1. The rectangle frames the join graph of the SPARQL query. The vertex with name (?article rdf: type bench:Article) means a triple vertex. The edge connects a join vertex with a triple vertex means the triple vertex needs to be joined with other vertices that also have edges with the same join vertex. Edges between triple vertices represent they have common variables. For example, there is an edge between vertex (?article rdf: type bench:Article) and vertex (?article dc:creator ?person), because both of them have variable ?article. The weight of edge represents the number of common variables between the two vertices. Therefore, the weight between (?article rdf: type bench:Article) and (?article dc:creator ?person) is 1. Without loss of generality, we limit ourselves to the equi-join as the most important representative of a join.

2.2 Index Structures

We use the open source RDF storage system Sesame [9] as RDF database backend that employs B-trees for RDF storage. RDF statements are organized in the B-tree structures. Currently, there is a *spoc* index and *posc* index in Sesame. These indexes are considered as B-trees that store a series of node identifiers in the index order. In Table 1, we will see all combinations of subject, predicate and object as well as the index structures that can be used to lookup them.

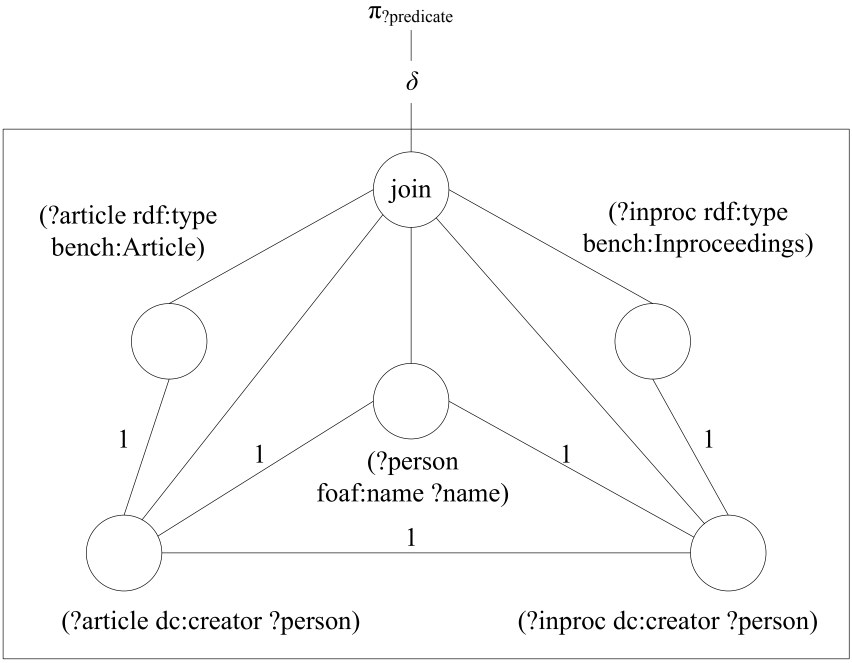

π?predicate
δ
join
(?article rdf:type bench:Article)
(?inproc rdf:type bench:Inproceedings)
1
1
(?person foaf:name ?name)
1
1
1
(?article dc:creator ?person)
(?inproc dc:creator ?person)

Fig. 2. The join graph of SPARQL query Q1

Table 1. All subject, predicate, object combinations and the corresponding indexes

Number of variable	Triple patterns	Index structures
0	(s p o)	*spo*
1	(?s p o)	*pos*
	(s ?p o)	*osp*
	(s p ?o)	*spo*
2	(?s ?p o)	*osp*
	(?s p ?o)	*pos*
	(s ?p ?o)	*spo*
3	(?s ?p ?o)	*spo, ops, pos*

B-tree index structure is prefix-matching. For instance, (s ?p o) triple pattern can be retrieved using the *sop* index or the *osp* index, and (?s ?p o) can be retrieved using the *osp* index. Thus, we can use the *osp* index to lookup both (s ?p o) and (?s ?p o).

Overall, we use the *spo* index to lookup (s p o), (s p ?o), (s ?p ?o) and (?s ?p ?o), use the *pos* index to lookup (?s p o) and (?s ?p o), and use the *ops* index to lookup (s ?p o) and (?s ?p o).

3 The ROS Algorithm

The ROS algorithm interleaves optimization and execution, and searches an optimal join order from the entire search space of possible excution orders. As soon as a join order is found to be superior than others, it will be executed.

First we will define some notations. Given a join graph $G = (V, E)$, a vertex $v \in V$, and an edge $e \in E$, $t(v)$ represents a table with all statements that satisfy the triple pattern, *card*(v) represents the number of statements that satisfy the triple pattern, *edges*(v) represents all triple pattern that have common variables with v, $w(e)$ represents the number of common variables between vertices that e collects, and *exec*(v, e) represents an excution join operation associated with e.

The detailed ROS algorithm is shown in Algorithm 1. It consists of two phases. The first phase initializes the join graph. The second phase alternates search space, exploration until all edges have been executed.

Phase 1 (line 1-4). For each vertex v, we query the RDF document, store the query results in $t(v)$, count the number and record it in *card*(v). For each edge e, we count the number of common variables between its two vertices, and store it in $w(e)$.

Phase 2 (line 5-25). The second phase of algorithm alternates between exploring the search space to find the superior join pair and excuting the pair, updating the graph until there is only one vertex connected to join vertex. Firstly, we find the smallest *card*(v) as the starting vertex (line 9-12), because of the feature of the equi-join that the size of final results is surely not larger than the size of the smallest vertex in a join operation. Secondly, we find the vertex that has the most common variables with the starting vertex. If both vertices have the same number of common variables with the starting vertex, we choose the vertex that has the smallest *card*(v), because this can cut middle results as much as possible (line 13-21). ROS will join the pair of vertices it finds, use a *vertex* v'' to denote the join results (line 22), delete the two vertices and add v'' to the join graph (line 23), refresh the join graph (line 24-25) and repeat the loop until there is one vertex under the join vertex (line 5).

The ROS algorithm decreases intermediate results by finding the smallest *card*(v) as the starting vertex. This can largely cut results that absolutely will not appear in the final results. The ROS algorithm builds a join graph, and initializes its vertices and edges in a more efficient way. For example, it executes the triple pattern using more efficient B-tree indexes. Then, it connects the vertices that have common variables instead of connecting all the vertices to build a fully connected graph. After initialization, The ROS algorithm finds the smallest edge, which also aims at reducing intermediate results.

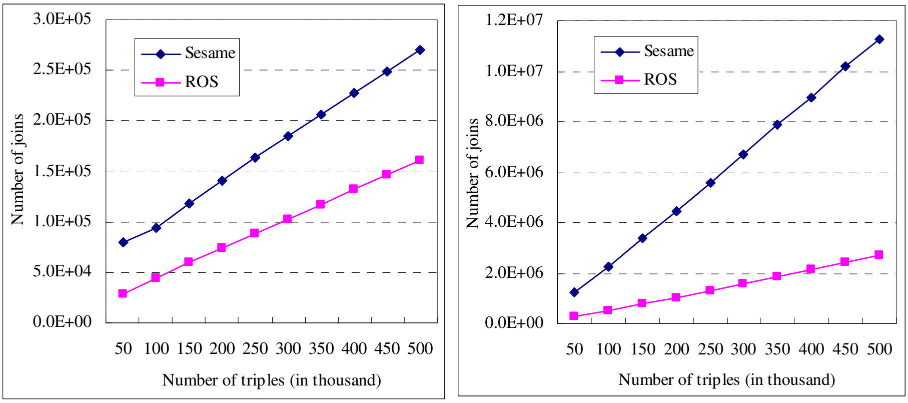
Algorithm 1. Run-time optimization of SPARQL queries

```
Input: Join graph $G = (V, E)$
for each $v \in V$ and $v$ is a triple pattern do
    $(t(v), card(v)) \in op(v)$;
for each $e = (v1, v2) \in E$ do
    $w(e) := countVar(e)$;
while there are more than one vertex to execute do
    vertex $v'$, $v1$;
    edge $e'$;
    $min$, $max$;
    for $v \in V \wedge edges(v) > 0$ do
        if $min > card(v)$ then
            $v' := v$;
            $min := card(v)$;
    for each edge $e = (v', v) \in edges(v')$ do
        if $max < w(e)$ then
            $v1 := v$;
            $e' := e$;
            $max := w(e)$;
        if $max = w(e)$ then
            if $card(v) < card(v1)$ then
                $v1 := v$;
                $e' := e$;
    $exec(v'', e')$;
    $updateGraph()$;
    for each $e \in edge(v'')$ do
        $w(e) := countVar(e)$;
```

4 Experiments

We choose Sesame-1.2.2 release as a platform for the implementation of ROS. We implement the ROS algorithm in Java. For all experiments here, we use a PC with 2.50 GHz Intel Core 2 Duo CPU, 3 GB memory, and running on a 7200 RPM disk with 500 GB capacity. Our experiments are based on SP^{2}Bench [10], which is a SPARQL performance benchmark. It is a meaningful analysis and comparison of both existing storage schemes for RDF data and evaluation approaches for SPARQL queries. It is a comprehensive and universal benchmark platform. It is settled in the DBLP scenario and comprised a data generator for creating arbitrarily large DBLP-like documents and a set of carefully designed benchmark queries.

We choose two representative SPARQL queries to shown the advantages of the ROS algorithm. Both of them are based on benchmark queries from SP^{2}Bench. The SPARQL queries Q1 and Q2 is shown in Fig. 1. Fig. 3 shows the experimental results of Q1 and Q2 between ROS and the optimization approach adopted by Sesame.



(a) Results of Q1

(b) Results of Q2

Fig. 3. Performance results of Q1 and Q2 between ROS and optimization algorithms adopted by Sesame

In Fig. 3, the *x*-coordinate shows the number of triples in SP^{2}Bench, and the *y*-coordinate shows the number of joins needed by doing a join operation. Fig. 2 shows that the join graph of Q1 has a relatively simple structure, which has only five vertices and five edges. In Fig. 3(a), we can see that with the growth of the number of triples, the number of joins ROS does is significantly reduced than Sesame optimization algorithms. Then, we can conclude that ROS does well in simple SPARQL queries.

Fig. 3(b) shows the performance results of SPARQL Q2 between ROS and the approach adopted by Sesame. It depicts that the number of joins in a join operation using ROS is lower than the approach Sesame adopted under the same condition. When the number of triples is 50 thousand, the number of joins ROS does is nearly half of the time used by Sesame. With the increasing in the number of triples, the advantage of ROS is more obvious. Thus, through this experiment, we can conclude that ROS also have the advantage in optimizing complex SPARQL queries.

ROS determines query plans in run time, instead of determining the query plans in compile phase. Doing this in run time can obtain good performance, since it reduces the size of the results and determines the best query plan. ROS uses intermediate results to refresh the join graph, and re-determine the best plan. In this way, it can fully consider the impact of the intermediate results on choosing the query plan, so the query plan is more accurate. ROS uses real values to select query plans, while the approach adopted by Sesame evaluates estimated values through B-trees to determine the query plan. Obviously, the approach that Sesame uses is inaccurate.

5 Conclusion and Future Work

In this paper, we have described the ROS algorithm, a run-time SPARQL optimization technology that intertwines query optimization into execution. The ROS algorithm does not need a stable cost model and can select a more accurate query plan. Furthermore, the use of join graph as part of an execution plan gives the ROS algorithm the possibility to handle a large class of SPARQL queries. Our experiments have revealed that ROS can select significantly better query execution plans than the existing query optimizers

Finally, we sketch future directions to extend and enhance the ROS algorithm. First of all, since we use real results of a triple pattern to evaluate the costs of a query plan, this always run into the risk of spending too much space for storing intermediate results. A future adaptation of ROS should use other methods to cut down results or find other ways to store results. Secondly, we plan to use a sampling technique to find an optimal path. Finally, we intend to study efficient ways of integrating operators such as sorting and duplicate-eliminating into join graphs.

Acknowledgments. This work was supported by the Program of 985 Project of Tianjin University, the National Science Foundation of China under grant number 61070202 and Seed Foundation of Tianjin University under grant number 60302010.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 147(5), 34–43 (2001)
2. Chaudhuri, S.: An Overview of Query Optimization in Relational Systems. In: 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 34–43. ACM, New York (1998)
3. Ioannidis, Y., Christodoulakis, S.: On the Propagation of Errors in the Size of Join Results. In: 17th ACM SIGMOD International Conference on Management of Data, pp. 268–277. ACM, New York (1991)
4. Haas, P.J., Swami, A.N.: Sequential Sampling Procedures for Query Size Estimation. ACM SIGMOD Record 21(2), 341–350 (1992)
5. Ling, Y., Sun, W.: A Supplement to Sampling Based Methods for Query Size Estimation in a Database System. ACM SIGMOD Record 21(4), 12–15 (1992)
6. Muralikrishna, M., DeWitt, D.J.: Equi-Depth Histograms for Estimating Selectivity Factors for Multi-dimensional Queries. ACM SIGMOD Record 17(3), 26–28 (1988)
7. Graefe, G., Ward, K.: Dynamic Query Evaluation Plans. In: 15th ACM SIGMOD International Conference on Management of Data, pp. 358–366. ACM, New York (1989)
8. Abdel Kader, R., Boncz, P., Manegold, S.: ROX: Run-time Optimization of XQueries. In: 35th ACM SIGMOD International Conference on Management of Data, pp. 615–626 (2009)
9. Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In: 1st International Semantic Web Conference, pp. 54–68 (2002)
10. Schmidt, M., Hornung, T., Lausen, G.: SP^{2}Bench: A SPARQL Performance Benchmark. In: 25th International Conference on Data Engineering, pp. 222–233 (2009)

The Research and Implementation of Heterogeneous Data Integration under Ontology Mapping Mechanism

Jing Bian[1,2], Hai Zhang[3], and Xinguang Peng[1,*]

[1] Colledge of Computer Science and Technology, Taiyuan University of Technology
[2] Department of Computer, Shanxi Medical College of Continuing Education
[3] Ministry of Information Technology Management
Shanxi Branch of Agricultural Bank of China
Taiyuan, China
Zhanghai3136937@163.com, {bianjing210,sxgrant}@126.com

Abstract. To deal with semantic heterogeneity among heterogeneous data sources, ontology is imported to the traditional data-integrated middleware. Domain ontology and local ontology are constructed as overall data view and local data view respectively. By establishing the mapping between domain ontology and local ontology, the problem of semantic heterogeneity is settled and semantic standard is reached. In this subject the way of automatic mapping is used to generate mapping relations innovatively. And an algorithm is given here. As a result, this way will provide an effective solution to the automatic integration of massive data.

Keywords: semantic heterogeneity, ontology, mapping, heterogeneous data, data integration.

1 Introduction

With the development of the network technology, information system leaded by computer application gradually enters electronic commerce, electronic government, and any other various domains. However, because of the obvious differences among the system hardware, operation system, databases selection and design in existing information systems, the datum generated by which are great heterogeneous. For a long time, the massive data sharing and uniform management of heterogeneous data sources have been delaying the steps of information construction in our country, and objectively formed the enormous obstacle to the ERP strategy of enterprises.

Analyzing the datum in relative domains, we found that they have the following heterogeneous features:

- Data model heterogeneity (such as the datum of structured SQL Sever database and semi-structured XML)
- DBMS heterogeneity (such as the Oracle and SQL Server)
- Semantic heterogeneity [1] (such as the phenomenon with the data sources about the same name, but the different meaning or the same meaning, but the different name, etc.)

[*] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 87–94, 2011.

The two former belongs to the heterogeneity at the grammatical stratum which have been basically solved by the methods such as using the corresponding query mode and connection driver. While semantic heterogeneity applies the mapping mechanism unified all the heterogeneous data sources. Considering to the advantages [2] of the sharing conceptual model expressing by ontology, this paper proposes a model based on ontology mapping mechanism. By constructing the domain ontology, extracting the local ontology and automatic mapping from local to domain to realize semantic integration of data heterogeneity.

2 The Design of System Architecture

The subject of heterogeneous data integration stated from last century, it has proposed successively relevant system architecture such as federated database, data warehouse, middleware etc. in view of the characters of heterogeneous data sources. However with the wide application of Web, diversification of integrated data sources and massive data, the heterogeneous integration system is changed day by day, especially the semantic heterogeneity problem. Study on combing ontology and middleware is becoming the trend of heterogeneous data integration fields. This model includes three stratums: user stratum, middleware stratum and bottom data sources [3]. Its architecture is showed in figure1.

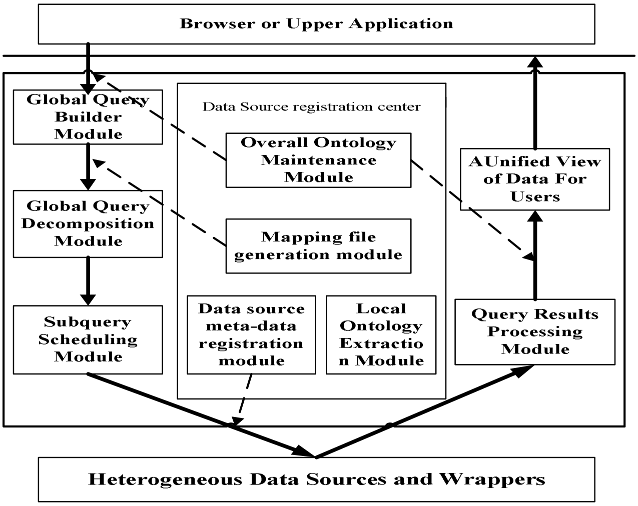

Browser or Upper Application
Global Query Builder Module
Data Source registration center
Overall Ontology Maintenance Module
AUnified View of Data For Users
Global Query Decomposition Module
Mapping file generation module
Subquery Scheduling Module
Data source meta-data registration module
Local Ontology Extractio n Module
Query Results Processing Module
Heterogeneous Data Sources and Wrappers

Fig. 1. System architecture diagram

The system operational flow is presented as follows:

- User stratum takes charge of user interaction and obtaining query parameters.
- Based on query parameters obtained by user stratum construct global query, which generates module by referencing domain ontology.

- Search decomposition module referencing mapping files from local ontology to domain one, decomposes the global query to sub-query. As a result the semantic of the sub-query is accordance with the semantic of the heterogeneous data sources [4].
- Sub-query scheduling module visit actual data source by a certain scheduling policy; execute the query language which is constructed by data source wrapper.
- Query result processing module executes the sub-query, returns result sets and processes data such as combine and delete.
- System returns user stratum last result for a unified structure.

Data source registry center include ontology mapping module and data source meta-data registration module. The former one is the core module among data source of semantic heterogeneity in all the system. And the latter one maintains the access parameters of data source, provide support to connect all actual data source.

3 The Core Module and Key Technology

This task realized integrated demonstrating sub-system on talent resources. Ontology mapping module which is the core of the whole system, has been designed to solve semantic heterogeneity among the data source in talents field. Based on the typical talents data source in Shanxi Province, we explain the key technology of the module as the follows.

3.1 Data Source Analyzing

At present most of the data sources (include some relational database SQL Server, Oracle, MySQL) are maintained by enterprises independently. Table 1 shows the typical data sources, among which data source 1 is Oracle data source, which is the information of agricultural talents in Shanxi province; data source 2 is SQL Server data source, which is the technology talents in Shanxi province; data source 3 is MySQL data source, which is the information of the agricultural experts.

From the above table we find that the problem of same semantic but variant name is obliviously. At the same time the data source applied different types of relation database. Moreover the real data source distributed in different network nodes maintained by different department independently. All of what brought difficulty to the uniform query of heterogeneous data.

3.2 Construct the Domain Ontology

For unified treatment of the heterogeneous data source, it is necessary to construct a uniform data module which could eliminate the semantic difference of the bottom data source. Considering the advantage of the accurate expression concept and the relationship within the concepts, we select the talent domain ontology to be the global data view for heterogeneous data source of this demonstration. During the construction processing, it uses the talents meta data of Chinese Academy of Science as a standard. The schematic diagram of talents domain ontology is given as follows:

Table 1. Experimental datasource filed name and Semantic Table

DataSource1 Field Name	DataSource1 Field Semantic	DataSource2 Field Name	DataSource2 Field Semantic	DataSource3 Field Name and Semantic
xm	Name	姓名	Name	Name
xb	Sex	性别	Sex	Sex
csny	Birthday	出生年月	Birthday	Birthday
gzdw	Workplace	工作单位	Workplace	Workplace
zc	Title	民族	Nation	Specialty
sxzy	Specialty	政治面貌	Party	Education
xl	Education	婚姻状况	Marriage	College
……	……	……	……	……

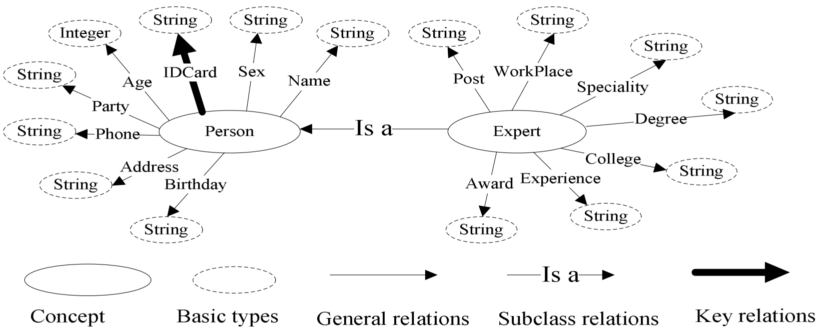


Fig. 2. Local diagram for talent domain ontology

To be a global view, the domain ontology has to describe most of (all in ideal condition) the conceptual semantic, of which terms have a certain normative in intra-industry for sharing the knowledge [5]. However, the information in real data source is often a subset of the standard domain ontology, which provides evidence for generating global query upward and reach the unified object of mapping with each local ontology downward.

3.3 Construct the Synonyms Lexicon

Considering the possibility of integrating mass data source, the system applies automatic mapping mechanism to construct the mapping relationship between domain ontology and local one. Whether there's a relation between domain ontology and local one in semantic synonyms databases is an important way to judge the mapping relation of both. Table 2 is the section of talent synonyms lexicon.

Table 2. Part of Synonyms thesaurus for talent domain

Overall concept	Local concept
Name	Xm，姓名，名字……
IDCard	Sfzh，身份证号，证件号码……
Age	Nl，年龄……
Address	Dz，地址，住址……
Workplace	Gzdw，工作单位……
Award	Shjl, 所获奖励, hjqk, 获奖情况……
Sex	Xb，性别……
……	……

In this table: The concept words constitute synonyms with themselves; it ignores the capitalized or lowercase problems of English letters and Chinese pingyin; Chinese words use fuzzy matching, for example, “work unit” has the same meaning with the “unit”. The synonyms lexicon supplies the service function String SearchSynonyms (String X), which could map within the concepts. Parameters X is a local concept, returns the synonymous character string of the global concept; returns “” if couldn’t find anything accord with conditions. In addition, synonyms lexicon guaranteed to completely establish the mapping relation by supporting dynamic update and extension.

3.4 The Ontology Automatic Mapping Algorithm

This is the main ontology automatic mapping algorithm flow:

```
Inpute: local ontology OWL files
1. load domain ontology
2. extract the related concepts in local ontology to
calculate the concept number CountA
3. For i=1 to CountA
    1) invoke Service SearchSynonyms (concept No.i)
    2) if return “″, i++
       else
        a) establish mapping between concept No.i and the
        character string.
        b) extract the concepts No.i in local ontology to
        calculate the CountB
       c) for j=1 to CountB
         1. invoke function SearchSynonyms (concept No. i,
         relation No.j)
          2. if return “″, j++
            else establish mapping between concept No.i
          and relation No.j, j++
     d) i++
output : mapping files
to suppose:
I: local ontology concept number
$X_i$: the relation number concept No.i of local ontoloty
M: domain concept number in synonyms lexicon
$Y_m$: synonyms number for the No.m domain concept
corresponding to local concept in synonyms lexicon
K: relation number of local ontology concept No.i in
domain ontology
```

We can use (1) to calculate the time complexity:

$$\frac{I * \sum_{m=1}^{M} (Y_m + 1)}{4} + \frac{\sum_{i=1}^{I} \left[X_i * \sum_{k=1}^{K} (Y_k + 1) \right]}{4} \tag{1}$$

For the condition of the actual data source and talent domain ontology, most of times the values (I, Xi, Ym, K) are taken in one magnitude, and the value of M are taken in two. So it can be Deduced that the total time complexity of the algorithm is at 4 magnitude, which is Far less than the operation speed of common computer, which ensured the feasibility of this algorithm.

4 Experiment

4.1 Experiment Environment

For the transplanting respect of the platform, this system applies Java programming language, of which user stratum use Struts 2 open source framework to realize the MVC three-tier architecture. The ontology management module applies protégé [6] which is developed by Stanford University as the visualization ontology development tool, and OWL language as the description language recommended by W3C. The user stratum published on the Tomcat5.x Server and the middleware use Jboss4.x Server.

Table 3. Mapping relation

Overall concept	Orcale	SQLServer	MySQL
Name	xm	姓名	Name
Sex	xb	性别	Sex
Post	zc	技术职称	“”
Phone	“”	电话	“”
Workplace	gzdw	工作单位	Workplace
……	……	……	……

4.2 Experiment Results

In view of the three heterogeneous data source above, by mapping between local ontology and domain one we construct the mapping relation:

Statistic the automatic mapping rate of the concept of local ontology to domain ontology as follows:

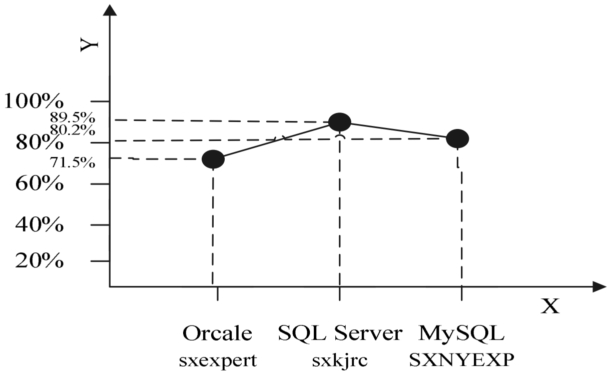


Fig. 3. Probability plot for automatically mapping

Table 4. System query results table

Name	Sex	PhoneNO.	……	DataSource
张维峰	男		……	Orcale
张名昌	男		……	Orcale
张喜文	男		……	MySql
张廷民	男	0351-2189538	……	MSSQL
张慧英	女	0351-3031235	……	MSSQL
……	……	……	……	……

The system support the fuzzy query based on the core concept of talents, for example the query based on name which include “张” in expert information. First the system construct the overall query: Select * From Expert Where Name=’%张%’

Then by referring the mapping files, it decomposes the overall query to the following sub query:

Oracle data source: Select * From sxexpert Where xm=’%张%’

SQL Server data source：Select * From sxkjrc Where 姓名=’%张%’

MySQL data source：Select * From SXNYEXP Where name=’%张%’

At last, it executes each of the sub query sentences and gets query results as follows:

The experiment proved that the demonstration system realized the unified search for talent mainstream relationship data source, of which the automatic mapping rate reached over 70% generally in integration processing. It is finally realized the goal of unification on data view and heterogeneous data sources, effectively solved the semantic heterogeneity problems existed among all kinds of data sources.

5 Conclusion

According to some prominent semantic heterogeneity problems today, we design and realize the heterogeneous data integration middleware under ontology mapping mechanism, which is on the basis of analyzing the present talents heterogeneous data sources. Now the demonstration system supported the mainstream relationship data source integration and realized the data source integration of Shanxi province talents part. In our next work, we plan to realize the dynamic loading of multi-domain ontology and the application of demonstration system to any other fields, meanwhile studying on the integration data source faced to semi-structured and structured.

Acknowledgment. This paper is supported by the Natural Science Foundation of Shanxi Province under Grant No. 2009011022-2 and Shanxi Scholarship Council of China Grant No. 2009-28.

References

1. Bellahsene, Z.: Data integration over the Web. Data and Knowledge Engineering. J. 265–266 (2003)
2. Ghoula, N., Khelif, K., Dieng-Kuntz, R.: Supporting Patent Mining by using Ontology-based Semantic Annotations. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, pp. 435–438 (2007)
3. James, J.L.: A data model for data integration. Electronic Notes in Theoretical Computer Science, 3–19 (2006)
4. Ives, Z., Florescu, D., Friedman, M.: An adaptive query execution system for data integration. ACM SIGMOD Int. Conf. on Management of Data, 299–310 (1999)
5. Savo, D.F., Lembo, D., Lenzerini, M., Poggi, A., RodríguezMuro, M., Romagnoli, V., Ruzzi, M., Stella, G.: MASTRO at work: Experiences on ontology-based data access. In: Proc.of DL 2010. CEUR, ceur-ws.org, vol. 573, pp. 20–31 (2010)
6. Yao, C., Shan, L., Hao., L.: Data Integration System Based on Ontology. Computer Engineering, 90–93 (2007)

Extracting Hyponymy Patterns in Tibetan Language to Enrich Minority Languages Knowledge Base

Lirong Qiu[1,2], Yu Weng[1,2], Xiaobing Zhao[1,2], and Xiaoyu Qiu[3]

[1] Information technology school, Minzu University of China, 100081 Beijing, China
[2] Minority Languages Branch, National Language Resource and Monitoring Research Center
[3] Institute of network and education technology, Shandong University of Traditional Chinese Medicine, Jinan, 250355
lirongqqq@163.com, mr.wengyu@gmail.com

Abstract. Semantic ontology is a formal, explicit specification of a shared conceptualization. The construction of semantic ontology knowledge base is the vital process in language processing, which is applied in information retrieval, information extraction and automatic translation. Hyponymy pattern is a basic semantic relationship between concepts, which is used to concepts acquisition to enrich ontology automatically. In this paper, the construction idea of multilingual ontology with unified criteria and interface are introduced, and hyponymy pattern is represented as a pair of a meaning frame defining the necessary information extraction in Tibetan language. The research of hyponymy relationship pattern can assist concept enrichment in ontology, which can reduce the cost during the ontology engineering process.

Keywords: knowledge base, semantic ontology, concepts acquisition, hyponymy relation.

1 Introduction

In early 90s of the 20th century, a lot of international symposia on ontology were held by the computer industry, ontology then became the hot topic of many artificial intelligence research groups, which include such branches as knowledge engineering, natural language processing and knowledge representation. The main reason for this trend is that, through ontology, the communication between people and people, people and machine, machine and machine can be built on the basis of shared knowledge [1].

China is a unified, multi-national country, with 56 nationalities in all. All these minorities have a total of 27 scripts of their own in current use, which are all computer-readable. In China, there are six languages to be commonly used, Chinese, Tibetan, Uygur, Mongolian, Kazak and Korean. For those languages, there are huge differences in gammar, pronunciation, spelling, vocabulary, which is increased the difficulty of interoperability between those languages with regard to information retrieval, information extraction and automatic translation.

However, use those languages also occurs some common features, and hold the common regularity to be complied. The common feature is the semantic meaning,

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 95–102, 2011.

which is specified when issue the word in the sentence. Those languages have some semantic properties in common, and the differences are shaped because of grammar, spelling etc.

Semantic ontology is a formal, explicit specification of a shared conceptualization [2]. The construction of semantic ontology knowledge base is the vital process in language processing, which is applied in information retrieval, information extraction and automatic translation.

The concept itself defined in the dictionary is not ambiguous; it can be associated with the real-world entity or object uniquely and accurately. However, in sentence processing, the concept of a word is closely related to the sentence. Let's take the word "Trojan horse" for an example; it can be interpreted into at least three meanings in these three sentences below:

(1) Trojan horse is a kind of toy.
(2) Trojan horse is a kind of sports equipment.
(3) Trojan horse is a kind of computer virus.

Therefore, the so-called ambiguity of concept is caused by polysemy, namely, a concept word with two or more different meanings. As in the case of Tibetan language, it can be translated into Chinese in different ways based on its context:

སློབ་ཕྲུག་རྣམས་ཚད་མ་སྦྱོང་བཞིན་འདུག་

(1) Students are learning Yin Ming Xue.

སྐྱེ་བོ་ཚད་མ་ནི་ཤཱཀྱ་ཐུབ་པ་ལྟ་བུའོ་

(2) The standard of saint like Shijiamoni

Furthermore, for Tibetan language, there exists a large number of foreign words and transliterated words. For example, the Chinese word "Chengdu" has different translations in the Tibetan language, such as ཁྲི་འདུ་ and གྲེང་ཏུ་.

The inherent fuzziness and ambiguity of semantics in language has made the work of machine analysis even more difficult. Word (binary data for the computer) is only a medium of semantics, and semantics is the core and critical part of communication.

For those people who got some knowledge, it is not hard to understand the specific meaning conveyed in the sentence based on the context. For example, if the word "Trojan horse" appeared in a text together with "computer" or "program", then it can be concluded based on common sense that "Trojan horse" here should mean the computer virus most likely.

The acquisition of hyponymy is a basic and vital problem in knowledge acquisition both from a computational linguistic perspective and from a theoretical linguistic one. Hyponymy is useful for the automatic creation or enrichment of an ontology, for tasks such as document indexing, information retrieval, question answering. On the other hand, these hyponymy patterns can be used in papers concerned with this semantic relation [3].

Given all that, the construction of semantic ontology knowledge base is the vital process in language processing, which is applied in information retrieval, information extraction and automatic translation. Hyponymy pattern is a basic semantic relationship between concepts, which is used to concepts acquisition in ontology automatically. In this paper, the construction idea of multi-language ontology with unified criteria and interface are introduced, and hyponymy pattern is represented as a pair of a meaning frame defining the necessary information extraction in Tibetan language. The research of hyponymy relationship pattern can assist concept enrichment in ontology, which can reduce the cost during the ontology engineering process.

2 Multilingual Knowledge Base

Semantic ontology is a formal, explicit specification of a shared conceptualization. To build ontology knowledge base for those minority languages is vital in language processing.

The minority language ontology knowledge base is fruitful for the information processing progress. Building multilingual ontology base is effective in relieving the condition of lack of language materials.

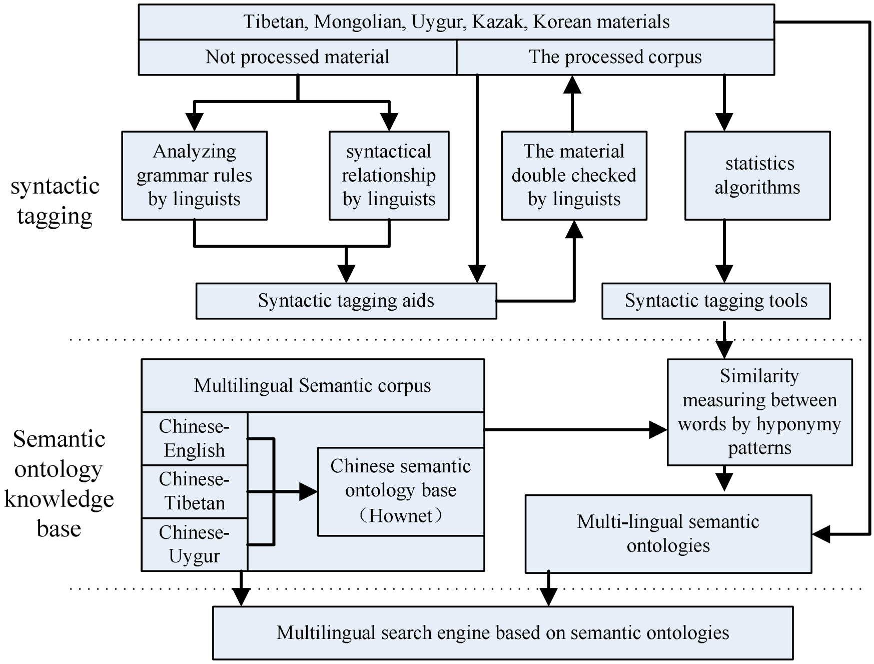

Tibetan, Mongolian, Uygur, Kazak, Korean materials
Not processed material
The processed corpus
syntactic tagging
Analyzing grammar rules by linguists
syntactical relationship by linguists
The material double checked by linguists
statistics algorithms
Syntactic tagging aids
Syntactic tagging tools
Semantic ontology knowledge base
Multilingual Semantic corpus
Chinese-English
Chinese-Tibetan
Chinese-Uygur
Chinese semantic ontology base （Hownet）
Similarity measuring between words by hyponymy patterns
Multi-lingual semantic ontologies
Multilingual search engine based on semantic ontologies

Fig. 1. The methodology of multilingual knowledge base construction

The methodology of multilingual knowledge base construction is shown in fig.1. Firstly, five minority languages materials which are not processed will be analyzed by linguists and will be syntactic tagged. The processed corpus will be double checked by linguists with the help of syntactic tagging tools. Secondly, the multilingual semantic ontologies will be built manually. During the creation of multilingual semantic ontology, the hyponymy pattern is represented as a pair of a meaning frame defining the necessary information, which can used to calculate the similarity between words.

The purpose of multilingual knowledge base construction is to produce a combination of dictionary and thesaurus that is more intuitively usable, to support automatic text analysis and artificial intelligence applications, and to realize great advantages in interoperability and invocation among multilingual languages.

Learning from the construction of HowNet, which is an ontology knowledge base in Chinese, the ontology structure of multilingual knowledge base is identified includes three parts:

(1) The basic properties, such as semantic code, the hyponymy relation, and the information of the word.

(2) The concept properties, such as using the core words to illustrate the meaning and collocation.

(3) The common grammar properties, such as subject, verb, object.

A semantic search engine is developed as a demonstration prototype system of multilingual ontology knowledge base. By now, the prototype system can be searched by words both in Chinese and in Tibetan. Also, when people input a keyword in Tibetan, the search engine can provide the results both in Chinese and in Tibetan, as long as the searching results are semantic related.

As far as we concerned, there hasn't been any research in minority languages processing of China on the level of semantic ontology.

3 Extracting Hyponymy Patterns in Tibetan Language

The acquisition of hyponymy is a basic and vital problem in knowledge acquisition, which is useful for the automatic creation or enrichment of ontologies and also can be used in extracting concepts concerned with this semantic relation.

The hypernym/hyponym relationships among the noun or verb synsets can be interpreted as specialization relations between conceptual categories.

There are three factors in hyponymy pattern extracting:

(1) The definition of hyponymy pattern in certain language;

(2) The selection of hyponymy pattern according to the sentence▯

(3) The algorithm to test the selected hyponymy relation.

In this section, the hyponymy patterns extracted in Tibetan language are introduced as an illustration of the multilingual ontology base.

3.1 Hyponymy Pattern Definition

We consulted the hyponymy definition of WordNet and gave the related definitions as follows:

Definition 1. Given a concept C_1 and C_2， the synonym set of C_1 is $\{C_1, C_1', \ldots\}$ and the synonym set of C_2 is $\{C_2, C_2', \ldots\}$. If the semantic meaning of C_1 was consumed by C_2, then C_1 and C_2 are the hyponymy relationship▯ in which C_1 is a hyponym of C_2, C_2 is called a hypernym of C_1, noted by $hr(C_1, C_2)$.

Definition 2. The hyponymy pattern space Ψ can be defined as a quad (*G, P, A, HR*).

G is the corpus, which are sentences in Tibetan language G={s_1, s_2, … , s_n}.

P is the hyponymic relations set P= {p_1, p_2, … , p_n} which is defined by human people and will be given in the next sub-section.

A is the algorithm set, which comprises the pattern learning algorithm.

HR is the hyponymy concepts set HR= {hr_1, hr_2, …, hr_n}, which can be learned by algorithms automatically.

3.2 The Hyponymy Patterns

Despite the significant amount of work done on acquiring hyponymy pattern automatically or semi-automatically recent years, we propose the hyponymic relation by human.

The reasons of getting hyponymy patterns in Tibetan manually are listed below: (1) Comparing with English or Chinese, not many available and frequently updated websites in minority languages can be downloaded. (2) As most minority languages, the information acquisition researchers face a great many of difficulties in dealing with being lack of text resource. (3) Further, the lack of the available minority languages electronic dictionaries and other useful ontologies acts as a brake to progress.

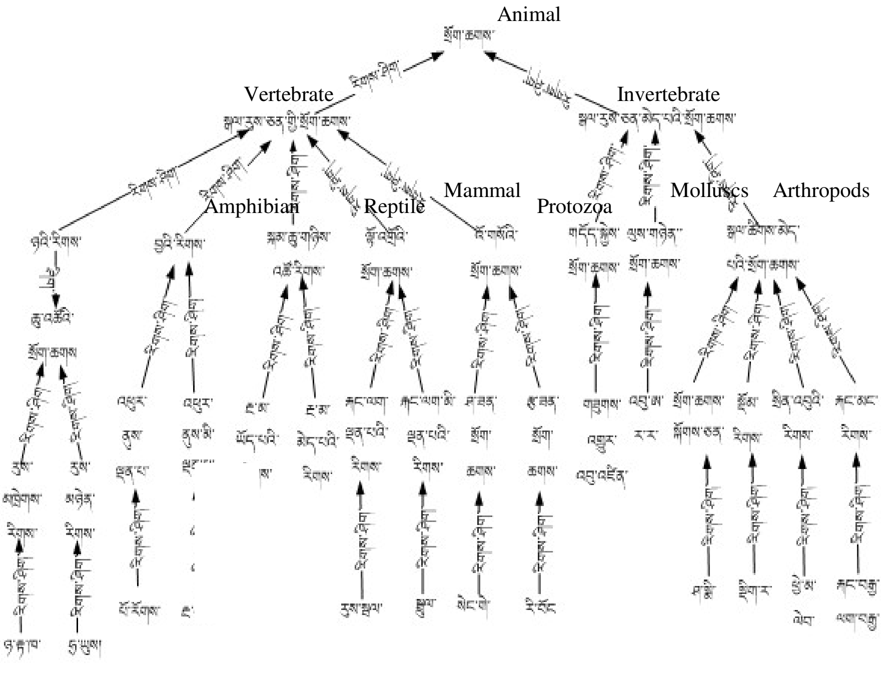


Fig. 2. The illustration of hyponymy concepts in Tibetan Language

The five hyponymy patterns in Tibetan language are defined as follows:

(1) One to one pattern

<?C1> 【ནི་】 <?C2> 【ཤིག་(ཞིག་/ཅིག་) ཡིན།】

(2) More to one pattern

<?C1> 【དང་】 <?C2>. 【 】 . <?Cm>. 【ལ་སོགས་པའི་】 .<?Cm+1>

(3) One to more pattern

<?C1>. 【ནི་】 <?C2>. 【ཤིག་(ཞིག་/ཅིག་)ཡིན་ཡང་】 .<?C3> 【ཤིག་(ཞིག་/ཅིག་) ཀྱང་ཡིན།】

(4) More to more pattern

<?C1>. 【དང】 .<?C2>. 【ནི་】 .<?C3>. 【ཤིག་(ཞིག་/ཅིག་)ཡིན་ཡང་】 <?C4> 【ཤིག་(ཞིག་/ཅིག་)ཀྱང་ཡིན།】

(5) Multiple level pattern

<?C1>. 【ནི་】 .<?C2>. 【ནང་ནས་】 .<?C3> 【ཤིག་(ཞིག་/ཅིག་)ཡིན།】

Figure 2 shows the illustration of hyponymy concepts in Tibetan Language, which is extracted from the texts.

3.3 Extracting Hyponymy Patterns

Given the hyponymy pattern set P={p_1, p_2, …, p_m}, and corpus *G*, there are sentence set S={s_1, s_2, …, s_n} in *G* and pattern p_1, p_2, …, p_k ($p_i \in P$, i=1, 2, … k). Then given $\forall$ s$\in$S and $\forall$ $p_i \in P$, if sentence s matches and p_i according to the pattern match algorithm, that can be noted by (s, {p_1, p_2, …, p_k }). If there are no pattern matches s, that can be noted (s, Ø).

For example, there is a sentence in the corpus:

གོས་སྒམ་ནང་ལ་སྟོད་ཐུང་དང་གོས་ཐུང་ ཕྱུ་པ་ལ་སོགས་པའི་གྱོན་ཆས་མང་པོ་འདུག།

There are some clothes in the wardrobe, such as jackets, trousers, gowns.

Pattern p can be defined as follows：

```
Defpattern hyponymy relation(one to more pattern)
{ Fundamental pattern:
<?C1> 【དང་】 <?C2>. 【 】 . <?Cm>. 【ལ་སོགས་པའི་】 .<?Cm+1>
The Hyponyms are <?C1>, <?C2> and <?Cm>
The hypernym are <?Cm+1>
}
```

The pattern match outcome is ：

གོས་སྒམ་ནང་ལ་སྟོད་ཐུང་ /དང་ /གོས་ཐུང་ / /ཕྱུ་པ་ /ལ་སོགས་པའི་ /གྱོན་ཆས་ /མང་པོ་འདུག།

Wardrobe /jackets /, /trousers /, /gowns/ clothes

Hyponyms：<?C1> = གོས་སྒམ་ནང་ལ་སྟོད་ཐུང་དང་གོས་ཐུང་

(*There are some jackets, trousers in wardrobe*)

Hyponyms：<?C2> = ཕྱུ་པ་ (*gowns*)

Hypernyms：<?C3> = གྱོན་ཆས་ (*clothes*)

The candidate hyponymy relation:

hr(སྟོད་ཐུང་དང་གོས་ཐུང་ , གྱོན་ཆས་) hr(*jackets\trousers, clothes*)

hr(ཕྱུ་པ་, གྱོན་ཆས་) hr(*gowns, clothes*)

The correct hyponymy relation:

hr(སྟོད་ཐུང་, གྱོན་ཆས་) hr(*jackets, clothes*)

hr(གོས་ཐུང་, གྱོན་ཆས་) hr(*trousers,clothes*)

hr(ཕྱུ་པ་, གྱོན་ཆས་) hr(*gowns, clothes*)

4 Related Work

As far as we know, there hasn’t been any research in minority languages processing on the level of semantic ontology.

Mr. Dong Zhendong, the creator of HowNet, has ever put it [4] that "natural language processing system will eventually need a more powerful knowledge base for support." The core of semantics is knowledge, and semantic ontology is a displayed formal specification of the shared conceptual model, which used to describe (specific areas) knowledge.

With its long history and splendid culture, the Tibetan language has long been the focus of many scholars and researchers both at home and abroad. Coincident with the development of digitalization, the information processing research in Tibetan has also gained rapid progress, covering such aspects as characters, words, phrases, sentences and chapters. The stage of sentence processing has also been addressed ambitiously, with such problems in the basic theoretical research as syntactic knowledge, semantic knowledge and pragmatic knowledge to be solved urgently [5].

Hyponymy is useful for the automatic creation or enrichment of ontologies, there are some research results in English and Chinese.

One of the first studies of hyponymy acquisition was done by Hearst [6]. Hearst proposed a method for retrieving concept relations from text by using predefined lexico-syntactic patterns. Other researchers developed other ways to obtain hyponymy, such as Brent proposes a method of syntactic information from text corpora by using verb sub-categorization frame recognition technique in [7]. Lei Liu proposes a method of extracting hyponymic relations from Chinese free text and using concept space to verify hyponymy in building a hyponymy lexicon in paper [8].

5 Conclusion and Future Work

China is a unified, multi-national country, with 56 nationalities in all. All these minorities have a total of 27 scripts of their own in current use, which are all computer-readable. Semantic ontology is a formal, explicit specification of a shared conceptualization. To build ontology knowledge base for those minority languages is vital in language processing, which can be applied in information retrieval, information extraction and automatic translation. Automatic acquisition and verification of hyponymy relations is a fundamental problem in knowledge acquisition.

In this paper, the construction idea of multilingual ontology with unified criteria and interface are introduced, and hyponymy pattern is represented as a pair of a meaning frame defining the necessary information extraction in Tibetan language. The research of hyponymy relationship pattern can assist concept enrichment in ontology, which can reduce the cost during the ontology engineering process.

The work of this paper is a part of our ongoing research work, which aims to provide an open reusable ontology knowledge base for further minority languages processing progress. Various experiments and applications have been conducting in our current research. Future work includes how to acquire and verify hyponymic relations from Tibetan free text, how to obtain sentences patterns automatically and how to verify the hyponymic relations with self features and context features.

Acknowledgments. Our work is supported by National 985 project "Development of science and technology innovation platform of information technology application in the minority national regions", the key projects in the National Science and Technology Pillar Program "Tibetan language and Uighur language monitoring technology project"(No.2009BAH41B04) and the "Science Fund for Youths" project of Minzu University of China(No. 1112KYQN39).

References

1. Neches, R., Fikes, R.E., Cruber, T.R., et al.: Enabling Technology for Knowledge Sharing. AI Magazine 12(3), 36–56 (1991)
2. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching very large ontologies using the WWW. In: ECAI 2000 (2000)
3. Mititelu, V.B.: Hyponymy Patterns. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246, pp. 37–44. Springer, Heidelberg (2008)
4. HowNet, `http://www.keenage.com/`
5. Jiang, D., Long, C.: Research of characters in Tibetan language – character, sound, code, sorting, graph, and rules of Latin interoperability. Social Sciences Academic Press, Beijing (2010)
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, Nantes, France, pp. 539–545 (1992)
7. Brent, M.R.: Automatic acquisition of subcategorization frames from untagged, free-text corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (1991)
8. Liu, L., Zhang, S., Diao, L.H., Yan, S.Y., Cao, C.G.: Using Concept Space to Verify Hyponymy in Building a Hyponymy Lexicon. In: Deng, H., Wang, L., Wang, F.L., Lei, J. (eds.) AICI 2009. LNCS, vol. 5855, pp. 479–486. Springer, Heidelberg (2009)

Discovering Atypical Property Values for Object Searches on the Web

Tatsuya Fujisaka[1], Takayuki Yumoto[2], and Kazutoshi Sumiya[3]

[1] Graduate School of Human Science and Environment, University of Hyogo, Japan
[2] Graduate School of Engineering, University of Hyogo, Japan
[3] School of Human Science and Environment, University of Hyogo, Japan
nd10u025@stshse.u-hyogo.ac.jp, yumoto@eng.u-hyogo.ac.jp,
sumiya@shse.u-hyogo.ac.jp

Abstract. Conventional search engines are able to extract commonplace information by incorporating users' requests into their queries. Users perform niche requests when they want to obtain atypical objects or unique information. In these instances, it is difficult for users to expand their queries to match their niche requests. In this paper, we introduce a query suggestion method for finding objects that have atypical characteristics. Our method focuses on the property values of an object, and elicits atypical property values by using the relation between an object's name and a typical property value.

Keywords: Web search, Atypical Object Search, Atypical Property Values, Relation between Object Name and Property Value.

1 Introduction

Recently, it has become very easy for users to survey information on everyday topics by using conventional search engines such as Google, Yahoo, etc. A user can obtain information on a topic just by using the topic's appellation as a search term. In addition, some search engines have suggestion functions, such as Google Suggest for Google Web searches. These functions suggest search terms to the user based on the strings that the user inputs for their query. The search terms that are suggested are often useful for the user to obtain detailed information about their topic of interest.

On the one hand, users sometimes wish to obtain niche information such as that related to objects that possess atypical characteristics. For example, a user may want to know about a type of curry that has unusual characteristics when compared to variety of curries. In this case, the user cannot formulate concrete queries to obtain information about the curry. If a query is made using the term "atypical curry", conventional search engines extract only pages that contain this phrase - they are unable to return more specific information that may be of interest to the user on the topic.

In this paper, we propose a method for suggesting search terms to the user for finding atypical objects. Specifically, it is focused on the object and its property values. Our system outputs atypical property values based on the relation between an object's name and its typical property value. When a user inputs an object's name and

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 103–112, 2011.

a typical property value for their query, our method outputs a pair that consists of an object’s name and an atypical property value. Thus, users can obtain information about atypical objects by using these pairs as their queries.

In this paper, we use Section 2 to explain work that is related to our research. Section 3 is used to define atypical objects, and explain a method for discovering atypical property values in detail. We then Section 4 focuses on an experiment that we use to evaluate our method. Lastly, in Section 5, we make our conclusions and present suggestions for further work on this topic.

2 Related Work

In terms of research on discovering atypical objects or unexpected information previously unknown to the user on the Web, there are Aramaki et al.'s work [2], Hattori et al.'s work [4], Otsubo's work [7] and Tsukuda et al.'s work [9]. Aramaki et al. defined the unknown information as being a content hole and proposed a "content hole search" method. Hattori et al. proposed a method that entails carrying out searches for peculiar images by converting peculiar color-names in Web pages into color-features. In their research, they used the term "peculiar images" to describe images that users cannot imagine easily as standard images. Otsubo developed a Web system which is able to encounter unexpected information related to an input query. Tsukuda et al. proposed a method that extracts a typical or an atypical recipe by adding or deleting an ingredient.

In both their and our research, the search targets are not standard objects. However, they employed overall evaluation of an object or a topic, or a particular category of object such as color-features or cooking ingredients, whereas our method discovers atypical objects from a variety of properties of objects by inputting a user viewpoint.

In terms of research on finding a property and property values, there are Hattori et al.'s work [3] and Tsuruta et al.'s work [10]. Hattori et al. proposed a method to extract appearance description of objects such as animals and geographic features. They found properties of the objects by using an image search engine, and extracted property values related to the properties on the Web. Tsuruta et al. proposed a method that automatically extracts a pair of a property name and a property value as basic corporate property from official Web sites. They excerpted only parts of the page which are written a property name by using the HTML structure. Then, they composed a set of property names by analyzing the parts and detected property values by using elements of the set.

Both their and our research extracts property values. However, our method doesn't use explicit property names of objects because it is hard to appear on the Web. Due to this, we use a relation between an object name and property values instead of excerpting property names.

In terms of research on extracting a relation between two terms, there are Kato et al.'s work [5] and Ohshima et al.'s work [6]. Kato et al. proposed a method that searches for object names based on similarity of relations input by users. Particularly, they found terms which strongly connect two terms A and B by using difference between distributions of co-occurring terms. Ohshima et al. proposed a method for discovering related terms such as coordinate terms using Web search engines. In their research, bi-directional syntax patterns are used for Web search queries and for extracting related terms from Web search results.

Both their and our research extracts a relation between two terms. However, our methods have to consider finding a relation which is able to discover atypical objects. Due to this, we use a character string that occur between an object name and property values and evolve the objects.

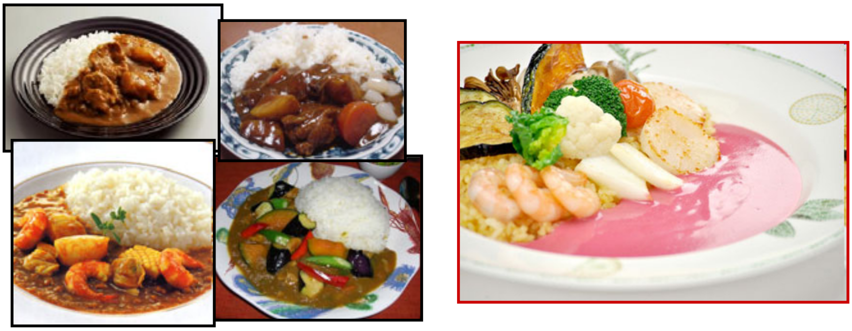
3 Our Approach

We define here atypical objects and explain our method.

3.1 Definition of Atypical Objects

We define an atypical object by using a given set of objects. An arbitrary object has properties and inherent values. In addition, objects in a given set have common properties. Under such conditions, when most objects in the given set have the same property value, we define this as being a typical property value. Using this definition, we designate an atypical object as being an object whose the property values are not typical.

We will elaborate on this with the following example. Suppose that the term "curries" as shown in Fig. 1 (a) and (b) defines a set of objects. Most curries possess "ingredient", "spice level", "color of sauce", etc. as properties. "Chicken curry", "pork curry", "vegetable curry," and "seafood curry" are shown in Fig. 1 (a). When we focus on the property "color of curry sauce", most objects have "brown" or "bistre" as a property value. We can therefore say that "brown" and "bistre" are typical property values of the property "color of curry sauce". On the one hand, "pink curry" [1] as shown in Fig. 1 (b) has "pink" instead of "brown" and "bistre" as its property value, and this is not a typical property value. Therefore, "pink curry" is an atypical object.



(a) A typical property value

(b) An atypical property value

Fig. 1. Typical or atypical property value for the property "color of curry sauce" of a curry

3.2 Finding the Atypical Property Values of an Object from a Typical Property Value

Our method is used to discover objects whose typical property value has been replaced with an atypical property value. For this purpose, there are two approaches to identifying atypical property values, which we describe as:

-We use only an object name as a query.
-We use an object name and a typical property value as a query.

For the first approach, it is easy for users to input a query but they cannot specify properties that they want to focus on. On the other hand, for the second approach, they can specify a property by using a property value. For example, a user can focus on "color of curry sauce", and if they input "curry + brown (or green)" to find a target, a system can output objects that have atypical property values that correspond to the user's intention. Thus, we adopt the second approach.

We describe a method for finding atypical property values in Fig. 2. The method comprises mainly three parts:

1. Collecting alternative property values from a typical property value.
2. Collecting alternative property values from candidate property values.
3. Detecting atypical property values by filters.

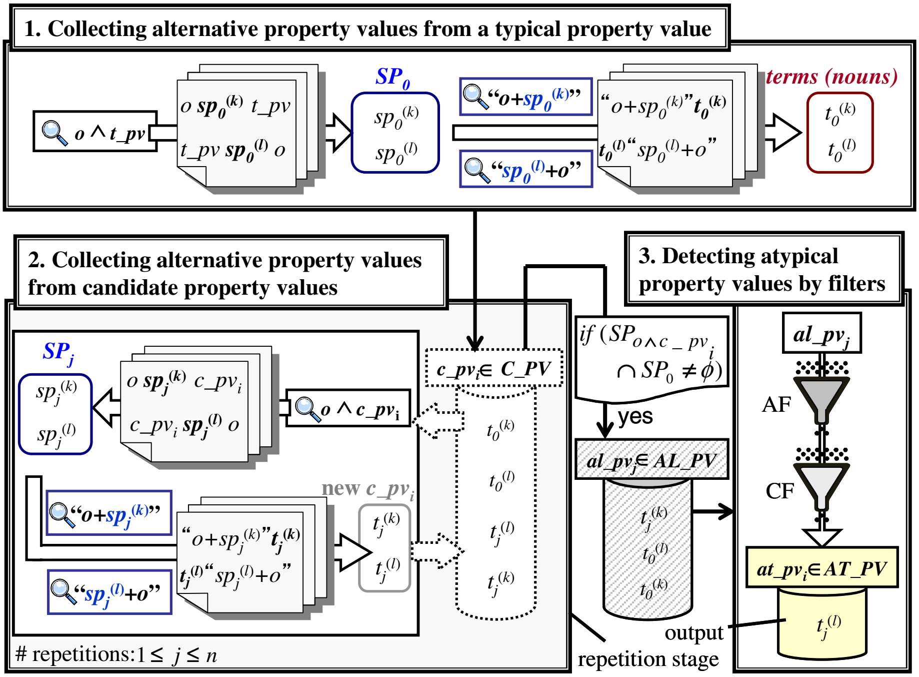


Fig. 2. A method for finding atypical property values

1. Collecting Alternative Property Values from a Typical Property Value. First, we extract the property from an object name and a typical property value. However, the property names do not always appear explicitly in Web pages. On the one hand, we can obtain a hypernym of a typical property value by using conceptual dictionaries such as WordNet [11] but it is not necessary to find property names.

We therefore focus on a character string between an object name and a typical property value instead of excerpting property names. For our research, we define a character string as being a syntax pattern and therefore the syntax patterns communicates a property.

For this, we excerpt syntax patterns that occur between an object name *o* and a typical property value *t_pv* as follows. First, we obtain Web pages from Web search results where the query was $In\_q(=o \wedge t\_pv)$ and then extract syntax patterns that exist between *o* and *t_pv* from the sentences containing both words. We then retrieve the syntax patterns $sp_0^{(k)}$ and $sp_0^{(l)}$ from o+$sp_0^{(k)}$+t_pv and t_pv+$sp_0^{(l)}$+o. We use only top *n* pages to reduce the processing time. SP_0 denotes a set of syntax patterns that were obtained by using *In_q*. The entirety of $sp_0^{(k)}$, $sp_0^{(l)} \in SP_0$ represents implicitly the same relation between *o* and *t_pv*.

Next, we explain about how to find candidate property values c_pv_i using SP_0. First, we gather top *n* Web pages of the Web search results by queries utilizing $sp_0^{(k)}$, $sp_0^{(l)} \in SP_0$. When the syntax pattern is $sp_0^{(k)}$, we obtain Web pages using the query "o+$sp_0^{(k)}$", and extract a term (noun) behind $sp_0^{(k)}$. On the one hand, when the syntax pattern is $sp_0^{(l)}$, we obtain Web pages using the query "$sp_0^{(l)}$+o", and extract a term (noun) before $sp_0^{(l)}$.

Third, we differentiate property values that can be replaced with a typical property value from candidate property values and call these alternative property values al_pv_j. For this, we regard all candidate values as being alternative property values.

2. Collecting Alternative Property Values from Candidate Property Values. In this part, we obtain more alternative property values from candidate property values. We use candidate property values instead of a typical property value.

First, we excerpt syntax patterns from an object name *o* and the c_pv_i. We pose the query $Re\_q(=o \wedge c\_pv_i)$ from *o* and c_pv_i, and obtain top *n* Web pages. Then, we extract syntax patterns $sp_j^{(k)}$ and $sp_j^{(l)}$ that occur between *o* and c_pv_i from sentences that contain both words. SP_j denotes a set of syntax patterns that were obtained by using *Re_q*. Second, we extract new candidate property values. We pose the query for a phrase search from *o* and SP_j that were excerpted by *Re_q*, and obtain top *n* Web pages. When the syntax pattern is $sp_j^{(k)}$, we obtain Web pages using the query "o+ $sp_j^{(k)}$", and extract a term behind $sp_j^{(k)}$. When the syntax pattern is $sp_j^{(l)}$, we obtain Web pages using the query "$sp_j^{(l)}$+o", and extract a term before $sp_j^{(l)}$. Third, we regard the term as being a new c_pv_i and detect alternative property values al_pv_j among the new c_pv_i. If $SP_{o \wedge \text{new } c\_pv_i} \cap SP_0 \neq \phi$, we regard the new c_pv_i as being the al_pv_j.

We repeat this process, comprising the extraction of syntax patterns from candidate property values and the extraction of candidate property values from syntax patterns, until we cannot find new candidate property values.

3. Detecting atypical property values by filters. Finally, we extract atypical property values at_pv_i only from alternative property values using an association filter (AF) and a co-occurrence filter (CF). These detect typical property values, and we regard the remaining property values as being at_pv_i.

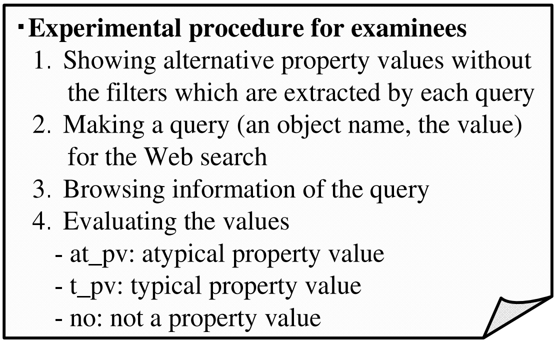
The association filter eliminates property values that are associated with an object name. Nouns in sentences that contain the object name are associated with the object name. We obtain top n Web pages with the query o only and we use sentences in these pages. On the one hand, the co-occurrence filter eliminates property values that co-occur frequently with the object name. This filter uses the number of Web search results. We denote $Hw(q)$ as the number of search results when a query is q. If $Hw(o, al\_pv_j) >= Hw\ (o, t\_pv)$, then al_pv_j is filtered out.

4 Experiment

We prepared some queries that comprised pairs of an object name and a typical property value as shown in Table 1, and then extracted alternative property values. Here, Japanese style pubs are places where alcoholic beverages are consumed usually in a private room, which was therefore chosen as the query. We asked six subjects to evaluate the extracted alternative property values. The experimental procedure is illustrated in Fig. 3.

Table 1. Keyword pairs as queries

object name	typical property value
soft serve ice cream	chocolate
Japanese style pub	private room
chair	wood



▪Experimental procedure for examinees

1. Showing alternative property values without the filters which are extracted by each query
2. Making a query (an object name, the value) for the Web search
3. Browsing information of the query
4. Evaluating the values
   - at_pv: atypical property value
   - t_pv: typical property value
   - no: not a property value

Fig. 3. The experimental procedure for subjects

Table 2. Results for the evaluation experiment (precision)

query	precision	part of property values which are selected by many examinees as "at_pv"
soft serve ice cream, chocolate	0.45	{rice cracker, Japanese horseradish, kelp}
Japanese style pub, private room	0.45	{prison, mansion, end of Edo Period}
chair, wood	0.33	{paper, Styrofoam, cork}

First, we evaluated the extracted atypical property values based on precision using formula (1).

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{|AT\_PV_i \cap ANS_i|}{|AT\_PV_i|} \quad (1)$$

AT_PV_i is a set of atypical property values that were extracted by our method when a query was q_i, and ANS_i is a set of property values that an subject evaluated as being atypical property values. The *n* expresses the number of subjects.

In Table 2, we show the precision of each query. We found that our method could be used to extract atypical property values. For example, when the query was “soft serve ice cream, chocolate”, our method extracted "rice cracker". The cracker is a classical Japanese snack and it is not used as an ingredient, especially for sweets. Therefore, many subjects regard "rice cracker" in the ice cream as being an atypical property value.

To analyze our method in detail, we show the results that it generated in Table 3 and 5. In Table 3, we denoted the number of syntax patterns and candidate property values with the first and second parts of the method as shown in Fig. 3. On the method of the second part, we expressed the numbers in every loop. In other words, it corresponds to the number of inquiries on the Web. As shown in Table 3, the transit of all queries is a bell-shaped tendency and eventually converged after some loops.

In Table5, we presented the extracted property values for each query. The underlined property values were evaluated as being atypical by many subjects. In the column, we classified the property values that were found in the both parts. The property values in the row "before filtering" are all alternative property values. The property values in “before AF” were inputted into the association filter and italic property values were eliminated by it. The property values in “before CF” were inputted into the co-occurrence filter and italic property values were eliminated by it. Property values in “after both filters” were the property values that remained after applying both filters.

As shown in Table 5, we found many alternative property values with both parts for all of the queries. In particular, more alternative property values were not eliminated by using both filters for the second part of the method than for the first part. Therefore, the second part of the method is important and might include more atypical property values than the first part. In fact, the second part extracted more atypical property values than the first part for all queries.

Table 3. The number of syntax patterns and candidate property values with the both parts

query (o, t_pv)	the first part (|SP_0|, |C_PV|)	the second part (|SP_j|, |C_PV|)
(soft serve ice cream, chocolate)	(15, 84)	(27, 109)→(47, 99)→(51, 100)→(46, 83) →(23, 54)→(15, 14)→(6, 8)→(2, 1) 8 loops
(Japanese style pub, private room)	(15, 37)	(19, 74)→(34, 109)→(57, 95)→(67, 89) →(47, 56)→(22, 43)→(15, 12) 7 loops
(chair, wood)	(26, 129)	(77, 252)→(113, 200)→(148, 105) →(98, 302)→(45, 242)→(26, 37) 6 loops

Table 4. Evalutation of the filters

query	association filter	co-occurrence filter	both filter
soft serve ice cream, chocolate	(6+11)/36 = 47.2%	(3+9)/25 = 48.0%	(9+9)/36 = 50.0%
Japanese style pub, private room	(5+7)/21 = 57.2%	(3+5)/16 = 50.0%	(8+5)/21 = 62.0%
chair, wood	(4+5)/19 = 47.4%	(0+5)/15 = 33.3%	(4+5)/19 = 47.4%

However, the precision of all of the queries that are listed in Table 2 is on the whole low. In this study, our filters were used to eliminate typical property values or noise. The filters may not have eliminated part of the typical property values or an entire property value. In addition, the filters might also have eliminated atypical property values. Therefore, we evaluated the filters by examining the detection rate (*dr*) that is utilized for evaluation of a spam filter [8]. We define *dr* as follows:

$$dr = \frac{|TP| + |TN|}{|AL\_PV_i|} \times 100 \qquad (2)$$

AL_PV$_i$ is a set of alternative property values that was found by our method before applying our filters when a query was q_i. True Positive (*TP*) is a set of the property values that were eliminated by our filter and judged by subjects to be typical or noise. True Negative (*TN*) is a set of the property values that were not eliminated by our filter and that were judged by subjects to be atypical.

As shown in Table 4, the detection rates ranged from 47.2 to 57.2 for the association filter. On the other hand, the rates ranged from 33.3 to 50.0 for the co-occurrence filter. Furthermore, for both filters, the rates ranged from 47.4 to 62.0, and half of all property values could be detected.

However, the accuracy was not high when both filters were used. In particular, the detection of *TP* was bad. For the association filter, we regarded the nouns of sentences that contained the object name in the Web contents as being property values that were associated with an object. We believe that this hypothesis was too simplistic. On the

Table 5. Extracted words for the query of chair and wood

query: **(soft serve ice cream, chocolate)**		
phase	part 1 method	part 2 method
before filtering	bean paste, baked sweet potato, strawberry, yogurt, Chinese citron, pineapple, cherry, Japanese plum, melon, tomato, shrimp, highland, grape	soy sauce, Japanese horseradish, choice draft, loquat, kelp, rice cracker, honey, wine, gold foil, apple, cafe creme,coffee, vegetable, silk, alcohol, Japanese snack, rice wine, dough, plastic, ginger, soymilk, lemon, caramel
before AF	*bean paste* , baked sweet potato, strawberry, *yogurt* , Chinese citron, pineapple, cherry, *Japanese plum* , melon, *tomato* , *shrimp* , *highland* , *grape*	*soy sauce* , Japanese horseradish, choice draft, *loquat* , kelp, rice cracker, honey, wine, gold foil, apple, cafe creme, coffee, vegetable, silk, alcohol, Japanese snack, rice wine, dough, plastic, ginger, soymilk, *lemon* , *caramel*
before CF	baked sweet potato, *strawberry* , Chinese citron, pineapple, cherry, melon	Japanese horseradish, *choice draft* , kelp, rice cracker, honey, *wine* , gold foil, apple, cafe creme, coffee, vegetable, silk, *alcohol* , Japanese snack, rice wine, dough, plastic, ginger, *soymilk*
after both filters	baked sweet potato, Chinese citron, pineapple, cherry, melon	Japanese horseradish, kelp, rice cracker, honey, gold foil, apple, cafe creme, coffee, vegetable, silk, Japanese snack, rice wine, dough, plastic, ginger

query: **(Japanese style pub, private room)**		
phase	part 1 method	part 2 method
before filtering	Japanese-style,elementary school, established opinion, irori fireplace, enthusiasm, British style, homemade, Spain, the Showa Period, Indonesia, Germany, fireside, nationality	room of tatami, log house, end of Edo Period, veteran, mansion, prison, folk dwelling, tap
before AF	*Japanese-style* , elementary school, established opinion, irori fireplace, enthusiasm, British style, *homemade* , Spain, *the Showa Period* , Indonesia, Germany, fireside, *nationality*	room of tatami, log house, end of Edo Period, *veteran* , mansion, prison, folk dwelling, tap
before CF	*elementary school* , established opinion, irori fireplace, enthusiasm, British style, *Spain* , Indonesia, *Germany* , *fireside*	room of tatami, *log house* , end of Edo Period, mansion, prison, folk dwelling, tap
after both filters	established opinion, irori fireplace, enthusiasm, British style, Indonesia	room of tatami, end of Edo Period, mansion, prison, folk dwelling, tap

query: **(chair, wood)**		
phase	part 1 method	part 2 method
before filtering	cardboard, Chinese, stainless, paper, bamboo, resin, cane, Japanese cedar, Germany, metal, British, fiber, continent	leather, titanium, iron and steel, styrofoam, cork, OKAMURA
before AF	cardboard, Chinese, *stainless* , paper, bamboo, resin, *cane* , Japanese cedar, Germany, *metal* , British, fiber, continent	leather, titanium, iron and steel, styrofoam, cork, *OKAMURA*
before CF	cardboard, Chinese, paper, bamboo, resin, Japanese cedar, Germany, British, fiber, continent	leather, titanium, iron and steel, styrofoam, cork
after both filters	cardboard, Chinese, paper, bamboo, resin, Japanese cedar, Germany, British, fiber, continent	leather, titanium, iron and steel, styrofoam, cork

one hand, for the co-occurrence filter, we compared the number of Web search results for (o, al_pv_j) and (o, t_pv). If $Hw(o, al\_pv_j) >= Hw(o, t\_pv)$, al_pv_j was eliminated as a typical property value. In this work, we did not define an explicit threshold. For this, the number was very high when t_pv was too commonplace and the filter did not eliminate the al_pv_j because it was atypical. Thus, the detection rate was not high. Based on these results, we need to improve the detection method in the future for typical property values to resolve these problems.

5 Concluding Remarks

We proposed a method for finding objects that have atypical property values by utilizing query suggestions. We first defined target atypical objects and then focused on syntax patterns that exist between an object and a typical property value in Web pages. Here, the patterns correspond to a relation between two words. Finally, we obtained alternative property values by using these patterns, and from this we detected atypical property values. In the future, we will improve the filters ability to detect typical property values. In addition, we will perform comparative analysis among existing methods. Furthermore, we plan to further examine various objects, and create a robust model.

References

1. A Pink Curry, http://www.myspiritual.jp/2010/07/post-1731.html
2. Aramaki, E., Abekawa, T., Murakami, Y., Nadamoto, A.: Dialog analysis of the Community Type Content for Content Hole Search. Journal of DBSJ 7(1), 109–114 (2008)
3. Hattori, S., Tezuka, T.: Mining the web for appearance description. In: Wagner, R., Revell, N., Pernul, G. (eds.) DEXA 2007. LNCS, vol. 4653, pp. 790–800. Springer, Heidelberg (2007)
4. Hattori, S., Tanaka, K.: Search the Web for Peculiar Images by Converting Web-extracted Peculiar Color-Names into Color-Features. IPSJ Transaction on Database (TOD) 3(1), 49–63 (2010)
5. Kato, M., Oshima, H., Oyama, S., Tanaka, K.: Object Name Search from the Web Based on Relational Similarity. IPSJ Transaction on Database (TOD) 2(2), 110–125 (2009)
6. Ohshima, H., Tanaka, K.: High-speed Extraction of Related Terms by Bi-directional Syntax Patterns from Web Search Engines. DBSJ Journal 7(3), 1–6 (2008)
7. Otsubo, G.: Goromi-Web: Browsing for Unexpected Information on the Web. In: The 6th ACM SIGCHI Conference on Creativity & Cognition, pp. 267–268 (2007)
8. Spam Filiter, http://en.wikipedia.org/wiki/Type_I_and_type_II_errors
9. Tsukuda, K., Yamamoto, T., Nakamura, S., Tanaka, K.: Plus one or minus one: A method to browse from an object to another object by adding or deleting an element. In: Bringas, P.G., Hameurlain, A., Quirchmayr, G. (eds.) DEXA 2010. LNCS, vol. 6262, pp. 258–266. Springer, Heidelberg (2010)
10. Tsuruta, M., Masuyama, S.: How to Find and Extract Corporate Profile Information from Official Web Sites. IEICE Transactions on Information and Systems J94-D(6), 977–988 (2011)
11. WordNet, http://nlpwww.nict.go.jp/wn-ja/

An Indent Shape Based Approach for Web Lists Mining

Yanxu Zhu[1], Gang Yin[1], Huaimin Wang[1], Dianxi Shi[1], Xiang Li[1], and Lin Yuan[2]

[1] College of Computer Science and Technology
National University of Defense Technology, 410073 Changsha, Hunan, China
{zhu.yannick,jack.nudt}@gmail.com,
whm_w@163.com,{dxshi1998,shockleyle}@gmail.com
[2] College of Electronic Technology
Information Engineering University, 450004 Zhengzhou, Henan, China
fkefss@gmail.com

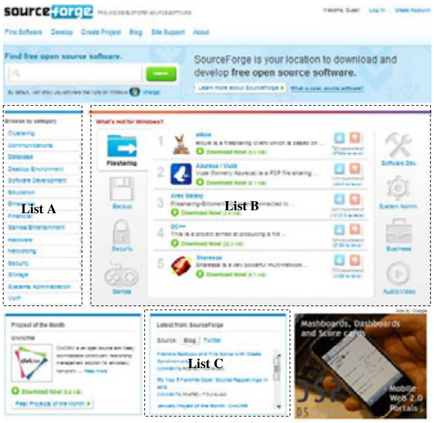
Abstract. Mining repeated patterns from HTML documents is a key step for typical applications of Web information extraction, which require efficient techniques of patterns mining to generate wrappers automatically. Existing approaches such as tree matching and string matching can detect repeated patterns with a high precision, but their efficiency is still a challenge. In this paper, we present a novel approach for Web lists mining based on the indent shape of HTML documents. Indent shape is a simplified abstraction of HTML documents in which tandem repeated waves indicate the potential repeated patterns to be detected. By identifying the tandem repeated waves efficiently with a horizontal line scanning along an indent shape, the repeated patterns in the documents can be recognized, from which the lists of the target Web page can be extracted. Extensive experiments show that our approach achieves better performance and efficiency compared with existing approaches.

Keywords: Repeated Patterns, Web Lists Mining, Indent Shape, Tandem Repeated Wave.

1 Introduction

Nowadays, Web data has become the largest public information source in the world. In a majority of cases, Web data is typically a description of objects retrieved from underlying databases and displayed in Web pages following some fixed page templates [15]. The various kinds of lists containing repeated data items are the most fundamental and common structures in page templates because of the intrinsic connection between repeated Web data items and relational database record lists. Figure 1 shows a Web page from sourceforge.net using three lists to organize its application data. The data item in each row usually holds valuable information for Web mining applications. Because rows in the same list are always generated by the same server-side HTML-generating component, their structures are usually similar or even totally identical. The formal expression of these row structures is called *repeated patterns*, which can also be treated as formal-defined templates such as regular expressions.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 113–121, 2011.



SourceForge is your location to download and develop free open source software.

Fig. 1. Lists region of a Web page. The figure only shows three lists as an example. Actually it contains more than three lists.

Repeated patterns mining is a hot topic in Web data mining area, and a lot of approaches have been proposed in the last decades [1]-[5]. String matching [2] and tree matching [3] are the two most important approaches. The time complexity of the two methods is at least O(N^2), where the problem size N is the length of the HTML encoded string and the number of nodes in the HTML DOM tree respectively. Because of its time-consuming nature, this may not be acceptable for most of Web mining applications. As a result, some heuristic methods were also proposed in recent years. Two typical examples are tag based approaches [6] and vision based approaches [4]. The former uses pre-defined tags as heuristic rules, but they cannot detect lists with starting tags beyond the scope of pre-defined tags. The latter uses visual contents as heuristic rules, but they have to request visual data from browser rendering engines, which introduces additional overhead.

In this paper, we introduce a novel **I**ndent **S**hape based approach for **L**ists **M**ining (**ISLM**) efficiently from Web pages. Given an HTML document following the *standard HTML indent style*, tandem repeated waves in its indent shape can almost capture all the features of its repeated patterns. We propose an efficient algorithm to find all the greatest tandem repeated waves in the indent shape, which can be used to identify all the lists easily according to the unique mapping between indent shapes and the lines in HTML documents.

2 Related Work

Repeated patterns mining of HTML documents are the basis for automatic Web data extraction. Since 1998, there are several automatic extraction systems have been proposed, such as IEPAD [2], MDR [3], Dela [8], and NET [9].

IEPAD [2] models an HTML document as a decoded binary string of the tag sequence and tries to find maximal repeated patterns using a PAT tree, which is similar to a suffix tree. MDR [3] uses DOM tree to find repeated patterns in cases where data records have a complex or nested structure, which is not addressed in

IEPAD. MDR searches multiple generalize-nodes using edit-distance similarity between sub trees. DeLa [8] and NET [9] concentrate on finding nested lists, which is based on tree traversal (not tree matching) and node comparison during the traversal. [7] uses trees to do the alignment instead of strings, which exploits nested tree structures to achieve much more accurate data extraction. However, all the above approaches does not consider time complexity in depth.

Many heuristic methods were proposed in order to improve performance. Diao et al [13] uses tags to identify boundary of list items, such as <div>, <P>, <TABLE> and <UL>. Lin et al [6] considers <TABLE> tags and uses an entropy based approach to discover informative ones. These methods are highly efficient but they do not work well when the lists are not distinguishable by the predefined tags. ViNTs [4] uses visual content features to identify boundary of the search result records, and then generates extraction rules combined with tag structure information. ViPER [10] identifies tandem repeats and visual context information for record segmentation, and enhances the extraction technique of MDR. VENTex [11] implements the information extraction from Web tables based on a variation of the CSS2 visual box model. ViDE [12] focuses on extracting regularly arranged data records and data items from deep Web pages. These methods mainly exploit visual features in the process of identifying data records or lists, hence they have to request visual data from browser rendering engines which introduce additional overhead.

This paper gains inspiration from the visual information based approaches. We treat an indent shape as a new kind of visual information for HTML documents, and indent shape can be easily retrieved from the source codes of HTML documents and will have little influence on performance. The differences between our approach and existing approaches include: (1) We use indent shape to model Web pages (now only HTML-based Web pages are considered) instead of DOM tree or encoded string, which leads to obvious performance improvement compared with existing approaches. (2) We only use starting tags of each line for similarity score calculation among list items instead of relying on a specific set of tags or identifying types of tags as previous approaches do.

3 Indent Shape Model

Generally, an HTML document is organized as a hierarchical structure consisting of layered elements, and each HTML element contains three parts [14]: a start tag, content, and an end tag, indicated by <tag> element content </tag>. Between each pair of tags, the element content may contain sub elements at lower layers. According to the above HTML coding standards, the following assumptions called *standard HTML indent style* are made in this paper:(1) The start tag of an HTML element begins with a new line and a line begins with an HTML tag;(2) The tag of a sub element at a lower layer starts with one additional indent unit (such as two blank characters) compared with their direct parent element;(3) The start tag and the end tag of the same element have same indents.

Note that some HTML documents do not follow the above assumptions. For example, HTML documents of Google (www.google.com.hk) have no line-break and are formatted into one line HTML code. In some other extreme samples, each line has no indents at all. But fortunately, lots of HTML format tools like Tidy(http://www.w3.org/People/Raggett/tidy) can be used to convert HTML documents into the ones satisfying the above three assumptions, and the transform operation only needs to traverse the document once with the complexity of O(N), where N is the total number of tags in that document. According to our preliminary experiments, more than 94.5% of 5000 HTML documents in real Web sites follow the three assumptions above.

For each line in a HTML document, the number of the blank characters before the first tag in this line is called its *indent distance*, and its starting tag (a start tag or end tag of an element) is called the *skeleton tag* of this line. Intuitionally, the combination of the line number of one line and its indent distance can be regarded as a point on a two-dimensional coordinate system, which is called the *tag point* of this line. Now we can give a formal definition of indent shape.

Definition 1, Indent Shape. Given an HTML document with n lines, its *indent shape* is an n-dimension vector $<(p_1, t_1), (p_2, t_2), \ldots, (p_n, t_n)>$, t_i is the skeleton tag of line i, and p_i is the tag point of line i and p_i is denoted by a 2-tuple (i, id_i) in which id_i is the indent distance of line i, $1 \le i \le n$. A zigzag curve with points p_1, p_2, …, p_n on a two-dimensional coordinate system, which is called an *indent shape curve* in this paper. The part located between any two tag nodes is defined as a *segment*.

Let w be a segment in indent shape, we use SLN(w) = i and ELN(w) = j to denote the start line number (SLN) and the end line number (ELN) of w respectively. If SLN(w')≤SLN(w) and ELN(w')≥ELN(w), w is called *sub-segment* of w', which is denoted by $w \subseteq w'$. If SLN(w') = ELN(w) +1 or SLN(w) = ELN(w') +1, w and $w'$are *contiguous*.

The similarity between the two sub-segments is calculated according to a classical method called Dice's coefficients. Also, there are many other methods, such as Jaccard's coefficient and cosine similarity etc. Given two sub-segments w_i and w_j of the same segment w, the vectors $st=<t_1, t_2,\ldots, t_k>$ and $st'=<t'_1, t'_2,\ldots, t'_k>$ are consist of distinct *skeleton tags*, the similarity of two vectors is calculated using formula (1) below:

$$\text{Dice}(st, st')=2*|st \cap st'|/(|st|+|st'|) \quad (1)$$

If the similarity is greater than or equal to the given threshold T, then it means w_i is similar with w_j, and the similarity equals to 1 means w_i is identical with w_j. The average value of each similarity between two contiguous sub-segments in w is greater than or equal to the given threshold T means w is a tandem repeated wave, otherwise it is not. This average value is defined with *self-similarity*.

Definition 2, Tandem Repeated Wave. Given an indent shape α and a segment w, there exists k sub-segments $w_i(i=1,2,\ldots,k)$,$w = w_1 w_2 \ldots w_k$, for each w_i (i=1,2,…,k-1) if the following two conditions are met: (i) w_i is similar with w_{i+1} or w_i is identical with w_{i+1};(ii) w_i and w_{i+1} are contiguous, then w is a tandem repeated wave.

Definition 3, Greatest Tandem Repeated Wave. Given a tandem repeated wave w in indent shape α, w is a greatest tandem repeated wave if and only if the following statements is true: for every tandem repeated wave w', if $w \subseteq w'$, then FLN(w)=FLN(w') and LLN(w)=LLN(w').

Based on the definitions above, segments corresponding to Web lists are tandem repeated waves, whose sub-segments are corresponding to records contained within the lists. The problem of mining Web lists is transformed to finding the entire greatest tandem repeated waves in the given indent shape.

4 Indent Shape Based Approach

This paper proposes a novel segmentation method by horizontally scan an indent shape with indent line. It has four basic steps:(1)divide indent shape into several segments.(2)calculate self-similarity among the segments.(3)detect tandem repeated waves based on self-similarity. (4) find greatest tandem repeated waves among these waves in (3).

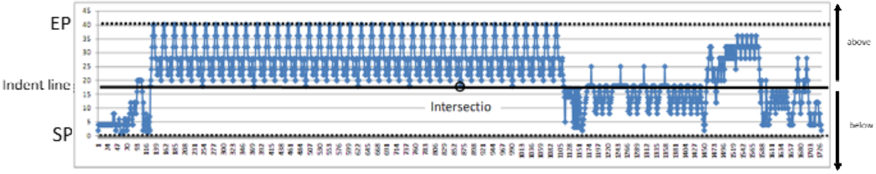
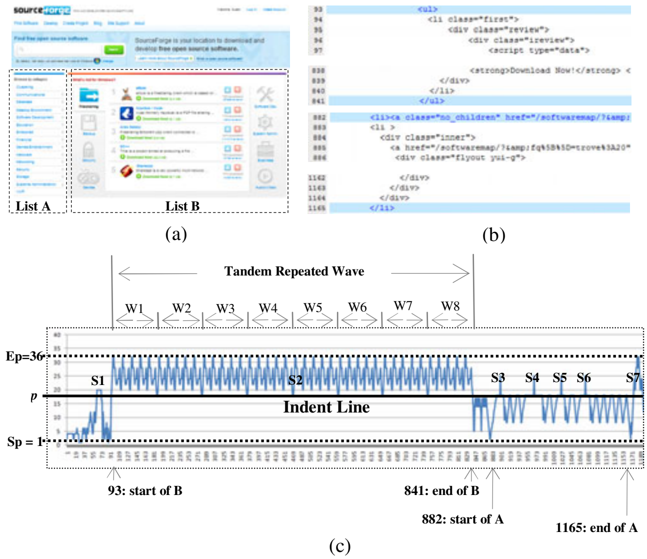


Fig. 2. The procedure of ISLM. *SP* stands for the start position, *EP* stands for the end position. *EP* and *SP* are shown with dotted lines, which are scanning boundary of intend line. The intend line is showed with solid line.

Figure 2 shows the procedure of ISLM (pseudo code is given in Appendix). Given an HTML document α and a threshold *T*, the indent shape of α is denoted by λ. We use an indent line to scan λ along the y-coordinate from the start position *SP* (the minimum indent in λ) to the end position *EP* (the maximum indent in λ) at different indent values occurred in λ. First, for each indent value in λ, α is divided into segments by the intersections with the indent line and λ, all the contiguous segments above or on the indent line are merged into one single segment selected as candidate segment; Then, for each candidate segment, its self-similarity will be calculated; Step three, if the self-similarity is greater than or equal to *T*, then the segment is identified as a tandem repeated wave; For the final step, all tandem repeated waves in λ will be discovered, among which containment relationship are checked. The greatest tandem repeated waves which are not contained by any of the others will be found.

In order to illustrate our approach, we give a case study. Figure 3(a) and 3(b) shows the home page of the Web site www.sourceforge.net and the indent contour of the different lines in its HTML document respectively. For each line in 3(b), its line number and the number of its indent characters (blanks) can be mapped to the x-coordinate and y-coordinate of a point in an indent shape, as shown in Figure 3(c) where *SP*=1 and *EP*=36. When the indent line moves to the position *p*=18, it will cut the indent shape into seven candidate segments among which there is one above *p* composed of eight contiguous sub-segments w_1, w_2, …, w_8. In this example, ISLM will identify two greatest tandem repeated waves. One corresponds to list B which starts at line 93 and ends at line 841, and the other corresponds to list A which starts at line 882 and ends at line 1165.



```
93                    <ul>
94                      <li class="first">
95                          <div class="review">
96                              <div class="ireview">
97                                  <script type="data">

838                                 <strong>Download Now!</strong> <
839                        </div>
840                      </li>
841                   </ul>

882          <li><a class="no_children" href="/softwaremap/?&amp;
883          <li >
884            <div class="inner">
885              <a href="/softwaremap/?&amp;fq%5B%5D=trove%3A20"
886               <div class="flyout yui-g">

1162              </div>
1163            </div>
1164          </div>
1165       </li>
```

Fig. 3. A Case study of ISLM. (a) is lists region of A and B in a Web page.(b) is HTML source code of the Web page, also it shows the indent contour of the different lines.(c) is the procedure and result of processing the Web page with ISLM. *p* is the current position of indent line. S1,S2,…,S7 are candidate segments. w_1, w_2, …, w_8 are sub-segments of tandem repeated wave S2.

5 Experiment

We compare our algorithm with existing methods MDR [3] and IEPAD [2] with two experiments using two data sets: data set 1 is from [5] and data set 2 is constructed from the home pages of practical Web sites, which are randomly selected from open source software (OSS) forges, whose pages mainly represent OSS project lists. The ground truth for data set 2 is manually constructed by experts of our research group. They identify all the lists contained in Web pages of data set 2, and compare the output results with the ground truth. We calculate the precision and recall by formula (2) and formula (3), where TP is the number of lists extracted from the pages correctly, FP (or FN) is the number of false positive (or false negative) lists extracted per page. Meanwhile we calculate F-score by formula (4). The algorithm in [5] is a successor of MDR and mainly focuses on nested lists, and we think MDR seems to be more convictive for this experiment.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{2}$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{3}$$

$$\text{F-score} = 2*\text{Precision}*\text{Recall}/(\text{Precision}+\text{Recall}) \tag{4}$$

Table 1. Performance results on the two data sets

Data Set	Data Set 1			Data Set 2		
Approach	IEPAD	MDR	ISLM	IEPAD	MDR	ISLM
Precision	67%	98.4%	98%	79%	96.4%	96.7%
Recall	39%	99%	100%	61%	96%	96.7%
F-score	0.49	0.99	0.98	0.69	0.96	0.967

Table 2. Efficiency results on data set 2

Page	Total characters	Total tags	Execute Time(ms)		
			IEPAD	MDR	ISLM
sourceforge.net	67624	3975	629154	2918	489
gitorious.org	11053	857	1433	109	9
appache.org	96876	4732	273	843	175
savannah.gnu.org	17358	832	400	48	21
gnome.org	17411	1206	1837	63	38
codeplex.search	105630	4624	919594	813	182
gna.project	15505	781	253	47	17
origo.project_list	23862	921	327	30	3
Average	44414.88	2241	194158.9	608.875	116.75

Table 3. Accuracy metrics changing with threshold *T*

	T<=0.3	**T=0.4**	**T=0.5**	**T=0.6**	**T=0.7**	**T=0.8**	**T=0.9**	**T=1.0**
Precision	83.7%	88.6%	**96.7%**	92.4%	95.5%	98.7%	100%	100%
Recall	99.6%	99.3%	**96.7%**	95.7%	36.7%	33.4%	5.6%	0.43%
F-score	0.910	0.93	**0.967**	0.940	0.530	0.500	0.106	0.008

The three algorithms are implemented with Java and tested on a PC (SONY VGN-NW18H, Intel Core CPU T6500 2.1GHZ*2, 4GB of memory, 32bit OS Win7). The threshold is 0.5 for IEPAD and 0.3 for MDR following the suggestions of their authors. Table 3 shows the parameter setting about threshold *T* and its affection to accuracy of our methods. According to table 3, when *T* = 0.5, all metrics of accuracy reach the maximum at the same time, the threshold for ISLM is 0.5. Table 1 shows the result of the first experiment: (1) Our approach is much better than IEPAD in all performance metrics with both data sets. (2) For data set 1, our approach has 100% recall and its precision and F-score are very close to MDR. For data set 2, our approach has a better performance compared with MDR.

Table 2 shows the result of the second experiment: our approach is significantly faster than IEPAD and MDR. Its average execution time is over 5 times faster than MDR and 1600 times faster than IPEAD.

6 Conclusion

Firstly, this paper proposes a novel indent shape model for HTML documents, based on which the problem of mining web lists is transformed to finding the entire greatest tandem repeated waves in the given indent shape. Secondly, the paper proposes a novel segmentation method by horizontally scan indent shape with indent line to find all the greatest tandem repeated waves. Finally, the paper carries out some extensive experiments, the results show that our approach achieves better performance and efficiency compared with existing approaches. More perfect performance of our approach is under research.

Acknowledgments. This paper is supported by National High-Tech Research and Development Plan of China under Grant (2007AA010301), National Natural Science Foundation of China under Grant (No.60903043), and the "Core electronic devices, high-end general chip and fundamental software" Major Project(2009ZX01043-001).

References

1. Embley, D.W., Jiang, Y., Ng, Y.K.: Record-Boundary Discovery in Web Documents. In: ACM SIGMOD International Conference on Management of Data, pp. 467–478 (1999)
2. Chang, C.-H., Lui, S.: IEPAD: Information Extraction Based on Pattern Discovery. In: The 10th International World Wide Web Conference, pp. 681–688 (2001)
3. Liu, B., Grossman, R., Zhai, Y.: Mining Data Records in Web Pages. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 601–606 (2003)
4. Zhao, H., Meng, W., Wu, Z., Raghavan, V., Yu, C.: Fully.: Automatic Wrapper Generation for Search Engines. In: The 14th International World Wide Web Conference, pp. 66–75 (2005)
5. Jindal, N., Liu, B.: A Generalized Tree Matching Algorithm Considering Nested Lists for Web Data Extraction. In: The SIAM International Conference on Data Mining, pp. 930–941 (2010)
6. Lin, S.-H., Ho, J.-M.: Discovering Informative Content Blocks from Web Documents. In: ACM SIGKDD, pp. 588–593 (2002)
7. Zhai, Y., Liu, B.: Web Data Extraction based on Partial Tree Alignment. In: The 14th International World Wide Web Conference, pp. 76–85 (2005)
8. Wang, J., Lochovsky, F.H.: Data Extraction and Label Assignment for Web Databases. In: The 12th International World Wide Web Conference, pp. 187–196 (2003)
9. Liu, B., Zhai, Y.: NET – A System for Extracting Web Data from Flat and Nested Data Records. In: Ngu, A.H.H., Kitsuregawa, M., Neuhold, E.J., Chung, J.-Y., Sheng, Q.Z. (eds.) WISE 2005. LNCS, vol. 3806, pp. 487–495. Springer, Heidelberg (2005)
10. Simon, K., Lausen, G.: ViPER: Augmenting Automatic Information Extraction with Visual Perceptions. In: Conference on Information and Knowledge Management, pp. 381–388 (2005)
11. Gatterbauer, W., Bohunsky, P., Herzog, M., Krpl, B., Pollak, B.: Towards Domain Independent Information Extraction from Web Tables. In: International World Wide Web Conference, pp. 71–80 (2007)
12. Liu, W., Meng, X., Meng, W.: ViDE: A Vision-Based Approach for Deep Web Data Extraction. IEEE Transactions on Knowledge and Data Engineering 22(3), 447–460 (2009)
13. Diao, Y., Lu, H., Chen, S., Tian, Z.: Toward Learning Based Web Query Processing. In: International Conference on Very Large Databases, pp. 317–328 (2000)
14. W3C, HTML 4.01 Specification (1999), `http://www.w3.org/TR/html401`
15. Liu, B.: Exploring Hyperlinks, Contents, and Usage Data. Springer, Heidelberg (2007)

Appendix: Pseudo Code of ISLM and Complexity Analysis

```
ISLM (Indent shape $\lambda$, Threshold T)
Output: set of greatest tandem repeated waves
for each distinct indent value occurred in $\lambda$
  get the intersections of the indent line and $\lambda$
     merge contiguous segments above or on the indent
    line into one segment, which inserted into set S
    for each segment Si of S
     if the number of sub-segments contained by Si $\geq 2$
       calculate self-similarity of Si
       if self-similarity of Si$\geq$ T
         insert Si into set of tandem repeated waves TRW
for each element of TRW
   if it is contained by any other element
    delete it from set
return TRW
```

The core idea of this algorithm is: tandem repeated waves capture all the repeated features of lists, these waves are different with other disorganized waves, so we can recognize lists by identifying tandem repeated waves. Line 4-5 divide indent shape into several segments, only contiguous segments above or crossed by indent line are corresponding to HTML elements. Line 6-10 calculate self-similarity of each segments, the self-similarity of which corresponding to the lists are more likely to be higher than the threshold value. Line 10 gets all the tandem repeated waves into the set TRW. Line 11-13 find all the greatest tandem repeated waves. After that we can extract all the lists easily. The complexity of ISLM is O($2kN_1$), where N_1 is the number of distinct skeleton tags of a given indent shape, k is the number of distinct indent values, in others words k is the depth of DOM tree of the standard HTML document. The complexity of converting HTML documents into the *standard HTML indent style* format is O(N_2), N_2 is the total number of tags of the HTML document. In the worst case, each tag is different which means N_1=N_2, the total complexity of our approach is O(KN_2),where K=2k+1.

Web Text Clustering with Dynamic Themes

Ping Ju Hung, Ping Yu Hsu, Ming Shien Cheng, and Chih Hao Wen

National Central University, Department of Business Administration,
No.300, Jhongda Rd., Jhongli City, Taoyuan County 32001, Taiwan (R.O.C.)
984401019@cc.ncu.edu.tw
Ming Chi University of Technology, Department of Industrial Engineering and Management
No.84, Gongzhuan Rd., Taishan Dist., New Taipei City 24301, Taiwan (R.O.C)
mscheng@mail.mcut.edu.tw

Abstract. Research of data mining has developed many technologies of filtering out useful information from vast data, documents clustering is one of the important technologies. There are two approaches of documents clustering, one is clustering with metadata of documents, and the other is clustering with content of documents. Most of previous clustering approaches with documents contents focused on the documents summary (summary of single or multiple files) and the words vector analysis of documents, found the few and important keywords to conduct documents clustering. In this study, we categorize hot commodity on the web then denominate them, in accordance with the web text (abstracts) of these hot commodity and their accessing times. Firstly, parsing Chinese web text of documents for hot commodity, applied the hierarchical agglomerative clustering algorithm--Ward method to analyze the properties of words into themes and decide the number s of themes. Secondly, adopting the Cross Collection Mixture Model which applied in Temporal Text Mining and the accessing times(the degree of user identification words) to collect dynamic themes, then gather stable words by probability distribution to be the vectors of documents clustering. Thirdly, estimate parameters with Expectation Maximization (EM) algorithm. Finally, apply K-means with extracted dynamic themes to be the features of documents clustering. This study proposes a novel approach of documents clustering and through a series of experiment, it is proven that the algorithm is effective and can improve the accuracy of clustering results.

Keywords: Documents Clustering, Temporal Text Mining, Extracting Theme.

1 Introduction

Due to the progress of network bandwidth and compression technology, the dissemination content of text, graphics, and sound has advanced to 2D and 3D animation and video. Through the Internet transmission, the computer not only enables messages to be received from all over the world, but also has become a favorite family entertainment medium. The well-known market research company Jupiter Research pointed out that global Internet advertising revenue in 2003 transcended that of Cable TV, making the Internet be the second-largest media after Wireless TV, while in relation to the rate of broadband Internet access (including ADSL and Cable Modem), Taiwan ranks second in the world after Korea. So, the

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 122–130, 2011.

competitive application of "online videos" by all industries and government departments is inevitable. The quality of playing online video will be enhanced along with the continual development of broadband network and compressed technology. From the experiences of Internet users, network media characteristics, and constant innovations, the form of digital video on the Internet will evolve to integrate more open and diverse thinking..

However, with the increased amount of Internet users, the challenge of website management is no longer how to obtain information from the vast amounts of data, but rather how to manage and classify the information. Because people generally lack sufficient time to analyze, digest, and absorb considerable information at a time, how to quickly and efficiently classify these huge amounts of online video with abstract (web text) and provide information to denominate category are the important issues for the website managers.

Recently, some scholars of text mining proposed temporal text mining (TTM)[5] analysis of time series documents, extracting the themes and analysis of themes evolution. This study adopted similar model, text description documents of online video collections would be arranged according to the time, extracting themes would be the features of clustering. Zhai et al.[6] divided these themes into three types: background themes, common themes and specific themes. In our studies, we would adopt the approach of TTM, and adapted these three kinds of themes to be the thesaurus of clustering groups. Our studies suggested that background themes can be regarded as stop words, the specific themes can be recognized as noise, and keep the common themes to be features of clustering.

This study is divided into five sections. Section 1 Introduction: explains the motivation, background, purpose, and scope of the study. Section 2 Related Work: review the related research, especially the TTM technology. Section 3 System Design: describes the research ideas, derives the study's proposed algorithms, and describes the system framework of our study. Section 4 Empirical Analysis: examines the cluster experiment conducted based on the research model, analyze the experimental results. Section 5 Conclusion and future research: presents the finding of our study, the contribution to the relevant fields, and the direction of future research derived in this study.

2 Related Work

The TTM refers to the text information collected in a period of time, and found the related mode of time. Since most of the text information has certain time stamps, much of TTM has been applied in a variety of fields, such as the news reports' summary of events and research trends revealed in scientific literature.

In 2002, Roy et al. [4] published an article on the trend detection method, especially trend detection for the initial development. They used the citation search of thesis inquiries to acquire citations to generate documents collection or the trend collection of a specific topic, which was then used to detect the initial emergence point.

In 2004, Morinaga and Yamanishi [3] published an article on mixture model to handle text streams and to develop the dynamic topic trend analysis. They proposed that with a unified thematic structure point of view, the weighted average probability model, through dynamic learning and tracking of text stream x, can establish the mixture model. Assuming $W = \{w_1,w_2,\ldots,w_d\}$ to be the term collection after all text streams have been treated through the word parsing and stop-words removal processing, *tf(w$_i$)* is the frequency or the occurrence times of the term w_i in a particular text stream *x*; *idf(w$_i$)* is the idf value of the term w_i ; and *tf- idf(w$_i$)* is the tf-idf value of term w_i, formula 1 represents the text stream x:

$$x = (\ tf(w_1), \ldots, tf(w_d)\ ) \text{ Or } x = (\ tf - idf(w_1), \ldots, tf - idf(w_d)\ ) \tag{1}$$

The author used the formula 2 to obtain probability $P(x\ /\ \theta{:}K)$ of the text stream x in the topic θ, and equals to the degree of distribution π_i of each topic θ_i (i = 1, ... , K) multiplied by the distribution probability $P(x\ /\ \theta_{i:}\ )$ of the text stream, where $\sum_{i=1}^{k} \pi_i = 1$. After that, we could identify the topic structure, detect the topic emergence and characterize the topic features.

$$P(x|\theta{:}K) = \sum_{i=1}^{k} \pi_i\, P(x|\theta_i) \tag{2}$$

Supposing each $P_i(x\ /\ \theta_{i:}\ )$ meet with the Gaussian density, the data dimension *d* of each text stream $P_i(x\ /\ \theta_{i:}\ )$ can be expressed as follows:

$$P_i(x|\theta_i) = \phi_i(x|\mu_i, \Sigma_i) \quad = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}\ (x - \mu_i)^T \Sigma_i^{-1}\ (x - \mu_i)\right) \tag{3}$$

Where μ_i is the real-valued vector, Σ_i is the *d* x *d* space matrix.

The TTM study in 2005 conducted by C. X Zhai and Q. Z Mei proposed the Cross Collection Mixture Model (CCMM) [5], which not only took into account the text structure of the clustering documents, but also used the background theme, common theme, and specific theme to deal with the text message of time stamps, while the background theme replaced the manual processing of stop-words removal, and the common theme was further used to facilitate the dynamic topic trend analysis.

3 System Design

The system of this study applied CCMM method based on the terms information of documents content, to calculate the correlation between terms, and further analyze the dynamic themes by identification of the degree of discrimination of terms information for documents clustering.

In this study, Chinese terms parsing process was adopted to analyze which terms w are in Content_title of each document d, and at the same time to inquire the number of times c(w, d) the terms w appeared in the document d. The practical operation adopted the term parsing process customer service program of Chinese terms system from the Chinese Knowledge and Information Processing (CKIP) of Academia Sinica for Chinese terms parsing. Usually the key words of documents would have some characteristics, because the verb and adjective have involved the application of vocabulary. When the subject under collection of a group of verbs and adjectives cannot be clearly expressed in its concept, this research intercepts the terms with characteristics of a noun as a dynamic theme for use, to facilitate nomenclature and description for the theme. Therefore, this study conducted term parsing process for the Content_title in all the documents, using terms beginning with N as filtering basis.

According to this research framework, the occurrence c(w,d) of all terms w_{ij} contained in document d_i is multiplied by the degree of discrimination in term identification $K_{d,i}$. The results are used to measure the distance between the documents for conducting clustering.

$$w_{ij} = K_{d,i} \times c(w,d) \tag{4}$$

The Ward's method of the Hierarchical method is adopted to cluster, while the Chebyshev distance is used to measure the distance between samples.

If the Chebyshev distance between two samples $d_i(w_{i1},w_{i2},..., w_{i|V|})$ and $d_j(w_{j1},w_{j2},..., w_{j|V|})$ is D, the formula is defined as follows:

$$D = max\left(| w_{i1} - w_{j1} |, | w_{i2} - w_{j2} |, \dots, | w_{i|V|} - w_{j|V|} |\right) \tag{5}$$

To judge by the agglomerative process, it is more appropriate to divide the documents content into Q clusters (Themes number = Q), and attribute the π_{dj} initial value $\pi^0_{di} = 1$ of each documents to the number i theme cluster; the rest is $\pi^0_{di} = 0$, in which $i \neq j$; $\sum_{j=1}^{Q} \pi^0_{d,j} = 1$ is expressed as follows:

$$\pi^0_{di} = \begin{cases} 1 & d \in \theta_i \\ 0 & d \notin \theta_i \end{cases}, i = 1,2,3,4 \tag{6}$$

The initial value of $P(w / \theta_j)$ is expressed as

$$P^0(w|\theta_j) = \lambda_C \times \frac{\sum_{d\in C_i} \pi_{d,j} K_{d,i}\, c(w,d)}{\sum_{d\in C_i} \sum_{w'\in V} \pi_{d,j} K_{d,i}\, c(w',d)} \tag{7}$$

The initial value of $P(w / \theta_{ji})$ is expressed as:

$$P^0(w|\theta_{ji}) = (1-\lambda_C) \times \frac{\sum_{d\in C_i} \pi_{d,j} K_{d,i}\, c(w,d)}{\sum_{d\in C_i} \sum_{w'\in V} \pi_{d,j} K_{d,i}\, c(w',d)} \tag{8}$$

This study adopted the CCMM by means of the probability distribution to conduct the term collection work, including the no. of θ_j(common theme), θ_{ji}(specific theme), and θ_B (background theme). The probability distribution for each term collection is expressed as follows:

$$P_d(w|C_i) = (1-\lambda_B)\sum_{j=1}^{Q}\{\pi_{d,j}[\lambda_c P(w|\theta_j) + (1-\lambda_C)P(w|\theta_{j,i})] + \lambda_B P(w|\theta_B)\} \tag{9}$$

The variable λ_B was set to gather terms in the ratio or weight of θ_B, while the variable λ_c was set as the model. The λ_c and λ_c is obtained through the researchers' past experiences in adopting the model; so the researchers themselves set the parameters. In order to generate CCMM, the probability distribution of all the terms was expressed by log-likelihood as follows:

$$\begin{aligned} log\, P(C) = \sum_{i=1}^{m} \sum_{d\in C_i} \sum_{w\in V} \{ c(w,d) K_{d,i} \log\{(1-\lambda_B)\sum_{j=1}^{Q}\{\pi_{d,j}[\lambda_c P(w|\theta_j) + \\ (1-\lambda_C)P(w|\theta_{j,i})] + \lambda_B P(w|\theta_B)\}\}\} \end{aligned} \tag{10}$$

This study, based on the EM algorithm, expresses the probability function that belongs to θ_B (background theme) of terms contained in all documents' content C={ $d_1, d_2,...,d_k$ } as follows

$$P(w|\theta_B) \quad = \quad \frac{\sum_{i=1}^{m} \sum_{d\in C_i} K_{\mathrm{d,i}} \times c(w,d)}{\sum_{i=1}^{m} \sum_{d\in C_i} \sum_{w'\in V} K_{\mathrm{d,i}} \times c(w',d)} \tag{11}$$

The values of $\lambda_{B\,,,}\lambda_c$ are to be set by the researchers themselves, and then using the Maximum Likelihood Estimate (MLE) [1] estimation method, the best parameter values are found for the log-likelihood

$$\boldsymbol{P(z_d, C_i, w = j)} \;=\; \frac{\pi_{d,j}^{(n)}[\,\lambda_C p^{(n)}(w|\theta_j)+(1-\lambda_c)p^{(n)}(w|\theta_{j,i})\,]}{\sum_{j'=1}^{Q} \pi_{d,j'}^{(n)}[\,\lambda_C p^{(n)}\left(w\middle|\theta_{j'}\right)+(1-\lambda_c)p^{(n)}\left(w\middle|\theta_{j',i}\right)]}$$

$$\boldsymbol{P\left(z_d, C_i, \theta_j, w = C\right)} \;=\; \frac{\lambda_C p^{(n)}(w|\theta_j)}{\lambda_C p^{(n)}(w|\theta_j)+(1-\lambda_c)p^{(n)}(w|\theta_{j,i})}$$

$$\boldsymbol{P(z_d, C_i, w = B)} = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B)+(1-\lambda_B)\sum_{j'=1}^{Q} \pi_{d,j'}^{(n)}[\,\lambda_C p^{(n)}\left(w\middle|\theta_{j'}\right)+(1-\lambda_c)p^{(n)}\left(w\middle|\theta_{j',i}\right)]}$$

$$\boldsymbol{\pi_{d,j}^{(n+1)}} \;=\; \frac{\sum_{w\in d} K_{\mathrm{d,i}}\, c(w,d)\, P(z_d,C_i,w=j)}{\sum_{j'=1}^{Q} \sum_{w\in d} K_{\mathrm{d,i}}\, c(w,d)\, P(z_d,C_i,w=j')}$$

$$\boldsymbol{P^{(n+1)}\left(w\middle|\theta_j\right)} \;=\; \frac{\sum_{i=1}^{m} \sum_{d\in C_{\mathrm{i}}} K_{\mathrm{d,i}}\, c(w,d)\, [1-P(z_d,C_i,w=B)\,]P(z_d,C_i,w=j)P\left(z_d,C_i,\theta_j,w=C\right)}{\sum_{w'=1}^{|V|} \sum_{i=1}^{m} \sum_{d\in C_{\mathrm{i}}} K_{\mathrm{d,i}}\, c(w,d)\, [1-P(z_d,C_i,w'=B)\,]P(z_d,C_i,w'=j)P\left(z_d,C_i,\theta_j,w'=C\right)}$$

$$\boldsymbol{P^{(n+1)}\left(w\middle|\theta_{j,i}\right)} = \frac{\sum_{i=1}^{m} \sum_{d\in C_{\mathrm{i}}} K_{\mathrm{d,i}}\, c(w,d)\, [1-P(z_d,C_i,w=B)\,]P(z_d,C_i,w=j)[1-P\left(z_d,C_i,\theta_j,w=C\right)]}{\sum_{w'=1}^{|V|} \sum_{i=1}^{m} \sum_{d\in C_{\mathrm{i}}} K_{\mathrm{d,i}}\, c(w,d)\, [1-P(z_d,C_i,w'=B)\,]P(z_d,C_i,w'=j)[1-P\left(z_d,C_i,\theta_j,w'=C\right)]} \tag{12}$$

This study adopts K-Mean [2] to cluster all documents according to their features $\boldsymbol{\pi_{d,j}^{(n+1)}}$, to analyze whether or not its clustering results are corresponding to the classification of the original documents. On implementing K-means, the k number of data points is randomly chosen firstly to be the center of clusters. Its center of mass m_i and the total offset distance E are formulated as follows:

$$m_i \quad = \quad \frac{\sum_{x\in Si} \vec{X}}{|S_i|} \tag{13}$$

$$E \quad = \quad \sum_{i=1}^{k} \sum_{x\in Si} \mid x - m_i \mid^2 \tag{14}$$

In which x represents a data point, m_i represents the mass center of the cluster S_i, $\mid S_i \mid$ represents the number of data points covered by cluster

The processing steps of this study shows in Figure.1

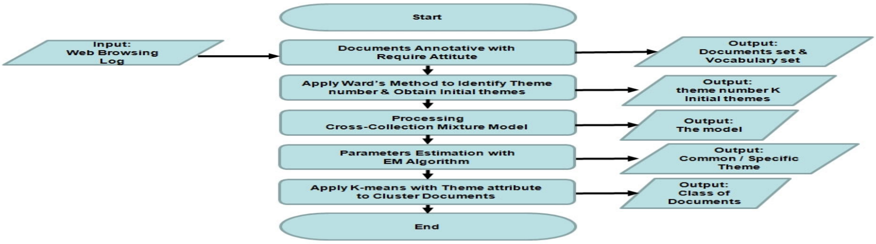


Fig. 1. System Design Process

4 Empirical Analysis

The data for this study were obtained from a domestic WAP online video website, of which the relevant data were created by the user in the daily hits. The website usually places the online video on the WAP page, when the user selects the online video, it immediately links the online video to the user's receiver to play, while the server records the video category, the time of hits, title content, transaction amount, and other relevant data of this transaction. This experiment collected according to the website's two categories of house keeper and female model all relevant data of transactions recorded in the database when users accessed that category by clicking a particular online video. Data used in this study were obtained from a year's record from January 2009 to December 2009. In the house keeper class, 848 cases of data were recorded, and in the female model class, 352 cases were recorded. But after exclusion of the terms without attributes of nouns, 823 cases of data were recorded in house keeper categories with of a total of 22,355 hits, and a total of 343 cases with 6,363 hits in the female model category. A total of 1166 cases of documents information can be expressed as C = { d_1, d_2,...,d_{1166} }. The terms data with attributes refers to the requirement for at least one word to be a noun, but not other terms attributes such as verb, adjective, expletive, prepositions, and so on. The information of documents filtered out in columns for analysis listed as follows:

1. Category: the original online video classification of documents came from the nomenclature type such as application type / content source / user types, etc
2. Time interval (m=4):the documents for one year is to be quarterly divided into four parts, C = $C_1 \cup C_2 \cup C_3 \cup C_4$,
3. Content_click ($K_{d,,i}$):The number of hits of document d in C_i is taken as the user's discrimination degree of identification on terms it contained, and the discrimination degrees of identification on all the terms the document d contains are equal
4. Content_title: the text information for narrative online video documents is composed of some of the terms of the collection V = { $W_1, W_2, \ldots, W_{|V|}$ }.

Totally 993 nouns were collected and the terms collection V = { $W_1, W_2, \ldots, W_{|V|}$ } and $|V|$ =993 established to conduct the following analysis of this study.

Based on Ward's method of grouping process, the clustering results are shown in Figure 2. It is judged more appropriate to divide the documents into 4 groups (Q=4), and to divide the documents into Cluster 1:637documents, Cluster 2: 76 documents, Cluster 3: 190 documents, and Cluster 4: 263 documents.

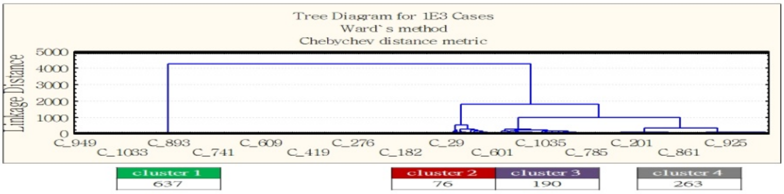
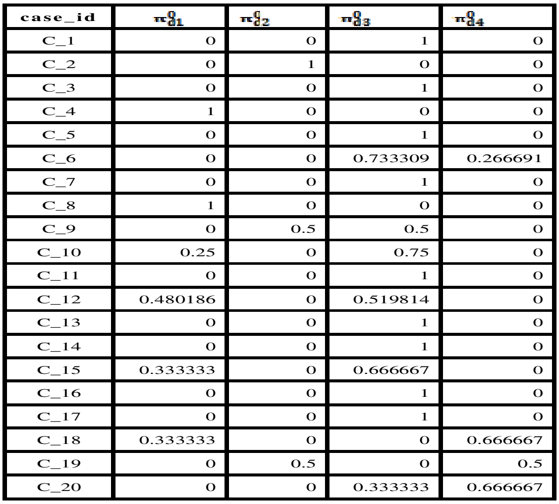


Fig. 2. Ward's method clustering results

The CCMM collects the terms in the way of probability distribution, including 4 sets of common theme θ_j, 4 x 4 sets of specific theme θ_{ji}, and background theme θ_B The λ_B, λ_C must be set firstly, and the value of its parameters are to be set by the researchers themselves. This study sets out the model to gather terms in θ_B in proportion or weight of $\lambda_B = 0.95$, because Content_title of the original data are mostly 18–25 words, the number of terms to be contained are small. The documents will be relatively verbose or mostly noisy, so it is essential to set larger λ_B value. The model was set to gather the terms in θ_j in the proportion or weight of λ_C, with the hope that the common theme can be more concise in order to facilitate the theme nomenclature and description.

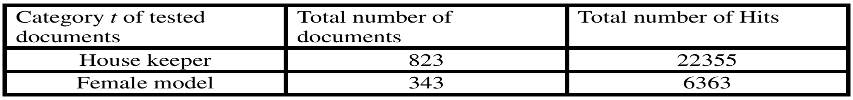
Table 1. The top 20 $\boldsymbol{\pi_{d,j}^{(n+1)}}$ and documents



case_id	π_{d1}^{0}	π_{d2}^{0}	π_{d3}^{0}	π_{d4}^{0}
C_1	0	0	1	0
C_2	0	1	0	0
C_3	0	0	1	0
C_4	1	0	0	0
C_5	0	0	1	0
C_6	0	0	0.733309	0.266691
C_7	0	0	1	0
C_8	1	0	0	0
C_9	0	0.5	0.5	0
C_10	0.25	0	0.75	0
C_11	0	0	1	0
C_12	0.480186	0	0.519814	0
C_13	0	0	1	0
C_14	0	0	1	0
C_15	0.333333	0	0.666667	0
C_16	0	0	1	0
C_17	0	0	1	0
C_18	0.333333	0	0	0.666667
C_19	0	0.5	0	0.5
C_20	0	0	0.333333	0.666667

This study adopted K-mean to cluster all documents in according with the themes features $\boldsymbol{\pi_{d,j}^{(n+1)}}$, to analyze whether or not the clustering results are corresponding to the classification of the original documents. A total of 1166 documents were set up in two clusters (K=2) to conduct the comparison with the original classification of the documents – totally 22,355 hits in 823 documents were recorded in the house keeper category, and totally 6,363 Hits in 343 documents in the female model category.

Table 2. Description of the experimental documents set



Category *t* of tested documents	Total number of documents	Total number of Hits
House keeper	823	22355
Female model	343	6363

Our study adopted the measure indicators F-measure to evaluate the clustering accuracy. F-measure is the harmonic mean for measuring the effect of the cluster.

$$F_{overall} = \frac{\sum_{t \in T} |t| \times F_t}{\sum_{t \in T} |t|} \tag{15}$$

F_{overall} is calculated to find out the average F-measure for all categories, which is representative of the overall clustering results and in which T stands for a collection of all documents (house keeper and female model), and ∣t∣ of traditional clustering method is the number of the specific category t within the cluster. ∣t∣ in this research model is the total number of hits of the specific category within the cluster.

Table 3. Comparative description of experimental results (3)

Test Documents Set	$F_{overall}$ (Traditional Clustering)	$F_{overall}$ (Our research Model)	Upgrading %
female model + house keeper	0.577967191	0.7262462	25.6553 %
	error rate (Traditional Clustering)	error rate (Our research Model)	Declining %
	0.430531732	0.299150359	30.5161 %

5 Conclusion and Future Research

The human mind is often filled with a wide range of innovative ideas while technological advancement is one of the important influences. Therefore, as the network technology continues to progress, the future application of audio-visual media in the Internet is expected to be richer and more varied in form. Therefore, how to rapidly and effectively compile information needed from the vast amount of data remains a very important issue. In the research field of data mining, many technologies have been developed to extract useful information from the vast amount of data, of which the documents clustering is considered one of the important technologies. In this study, based on the associations among the words provided in the Internet documents, the popularity or number of hits of users' identification of word information and the number of Q groups of the dynamic themes were analyzed to further provide documents clustering methods and the definitions and summary of the dynamic themes, which shall serve as a basis for documents cluster naming and group products or services consumers wish to find. This will in turn help the Internet management to quickly and accurately compile documents and information in response to consumers' needs and preferences.

In order to effectively improve the documents clustering quality, further research may strengthen the word association rule mining to identify co-occurring text phrases. Through the mining sequential patterns algorithm, the phrases in the text and the associations among the phrases have been found. The so-called "phrase" here refers to the sequential nature of the emergence of keywords instead of general phrases identified in grammar. The discovery of the associations among phrases is expected to enhance the quality of documents clustering.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EMalgorithm. Journal of Royal Statist. Soc. B 39, 1 (1977)
2. Khan, S., Ahmad, A.: Cluster Centre Initialization Algorithm for K-Means Clustering. Pattern Recognition 25, 1293–1302 (2004)
3. Morinaga, S., Yamanishi, K.: Tracking dynamics of topic trends using a _nite mixture model. In: Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 811–816 (2004)
4. Roy, S., Gevry, D., Pottenger, W.M.: Methodologies for trend detection in textual datamining. In: The Textmine 2002 Workshop, Second SIAM International Conference on Data Mining (2002)
5. Zhai, C.X., Mei, Q.Z.: Discovering Evolutionary Theme Patterns from Text-An Exploration of Temporal Text Mining. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 198–207 (2005)
6. Zhai, C.X., Velivelli, A., Yu, B.: A Cross-Collection Mixture Model for Comparative Text Mining. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, pp. 743–748 (2004)

Multi-aspect Blog Sentiment Analysis Based on LDA Topic Model and Hownet Lexicon*

Xianghua Fu, Guo Liu, Yanyan Guo, and Wubiao Guo

College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen Guangdong, 518060, China
fuxh@szu.edu.cn, {lgnagel7,choupigwb2006}@163.com,
guoyyszu@gmail.com,

Abstract. Blog is an important web2.0 application, which attracts many users to express their subjective reviews about financial events, political events and other objects. Usually a Blog page includes more than one theme. However the existing researches of multi-aspect sentiment analysis focus on the product reviews. In this paper, we propose a multi-aspect Chinese Blog sentiment analysis method based on LDA topic model and Hownet lexicon. At first, we use a Chinese Blog corpus to train a LDA topic model and identify the themes of this corpus. Then the LDA model which has been trained is used to segment the themes of Blog pages with paragraphs. After that the sentiment word tagging method based on Hownet is used to calculate the sentiment orientation of every Blog theme. So the sentiment orientation of the Blog pages can be represented by the sentiment orientation of multi-aspect Blog themes. The experiment results on SINA Blog dataset show our method not only gets good topic segments, but also improves the sentiment classification performance.

Keywords: Sentiment analysis, LDA topic model, Hownet lexicon.

1 Introduction

In recent years, there have been many interests in sentiment analysis research on text reviews. The existing sentiment analysis methods usually can be divided into two categories: (1) One type is based on part-of-speech (POS) tagging of words and sentiment lexicons. This type is proposed earliest by Peter [1] , and both Linhong Xu[2] and Weifu Du[3] also adopted this type methods for the sentiment analysis. Qun Liu[4] proposes establishing a believable vocabulary on semantic knowledge named Hownet, and then getting the sentiment polarity of words through comparison with the similarity between the words. Yanlan Zhu [5] succeeds in judging semantic orientation based on the Hownet. (2) The other type is based on machine learning. Such as Whitelaw uses support vector machine (SVM) to classify the sentiment orientation of movie reviews[6]. Weihao Lin proposes a learning method based on statistical model, which obtained that opinions reflected in the text through analyzing words[7]. Other machine learning methods also be used, such as Huifeng Tang adopts the SVM classification

* This work is supported by National Natural Science Foundation of China (60903114, 61003271, 61001185), Natural Science Foundation of Guangdong Province (7301329), Science and Technology Foundation of Shenzhen City (JC201005280463A).

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 131–138, 2011.

method [8], Xiangwen Liao makes use of the probability statistical model[9], Jun Xu uses the Naive Bayesian(NB) and maximum entropy[10].

With the development of the web2.0 technology, more and more popular users are willing to express their opinions on Blog. But few researchers pay attention to the sentiment analysis on multi-aspect Blog topic (topic also is called theme in this paper). In this paper, we propose a multi-aspect Blog sentiment analysis method based on the LDA topic model and the Hownet lexicon.

2 The Process of Multi-aspect Blog Sentiment Analysis

The process of multi-aspect Blog sentiment analysis is shown in Fig.1. At first we train LDA model with the Blog corpora and obtain some theme models. Then Kullback-Leibler (KL) divergence is used to calculate the similarity between the text paragraphs and theme models in order to determining the paragraph's theme. After that we find out the sentiment sentence in this Blog theme. Finally use the sentiment words tagging method to calculate the sentiment orientation of the sentences in each theme, and summarize the sentiment orientation of the Blog theme.

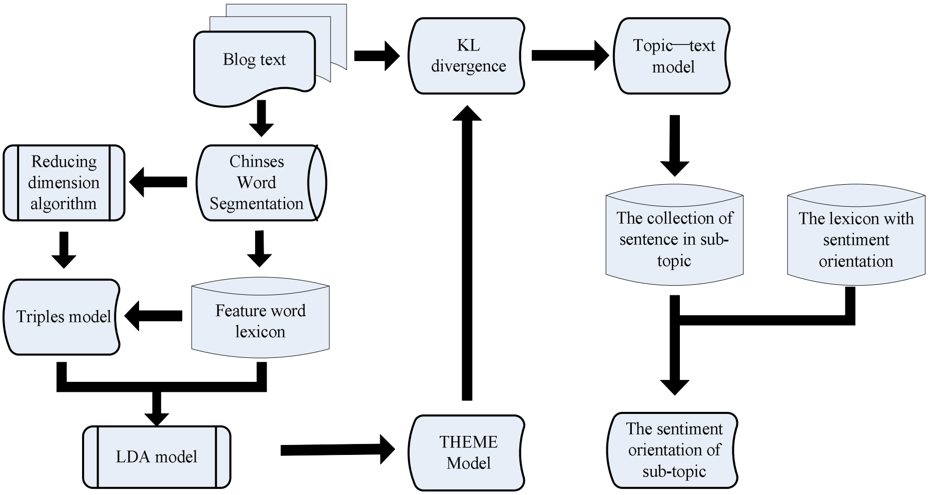

Blog text
KL divergence
Topic—text model
Reducing dimension algorithm
Chinses Word Segmentation
The collection of sentence in sub-topic
The lexicon with sentiment orientation
Triples model
Feature word lexicon
LDA model
THEME Model
The sentiment orientation of sub-topic

Fig. 1. The process of multi-aspect Blog sentiment analysis

3 Identify Blog Topic Based on LDA Model

3.1 Latent Dirichlet Allocation Model

Latent Dirichlet Allocation (LDA) [11] is a generative probabilistic model for the collections of discrete data, which is a three-level hierarchical Bayesian model. LDA based on the assumption that documents are mixtures of topics, where each topic is a probability distribution over words. The LDA's probabilistic graphical model is shown in Fig. 2.

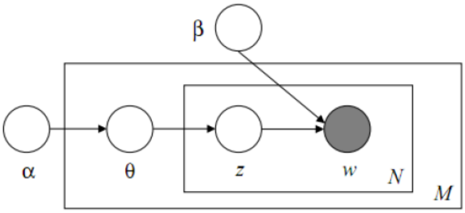


Fig. 2. The LDA's probabilistic graphic model

Let M is the total number of topics, N is the number of features words, β denotes a $M \times N$ matrix, and α is the super-parameter of prior distribution. w_{dn} denotes the feature word w which locates in the nth position of the document d . The topic is a distribution over words denoted by $z_d = \{w_{d1}, w_{d2}, ..., w_{dn}\}$, θ_d subjects the Dirichlet distribution $Dirichlet(\theta_d \mid \alpha)$. LDA model assumes the following generative process for each document $\mathbf{w}$:

(1) Choose $N \sim Poisson(\xi)$.

(2) Choose $\theta \sim Dir(\alpha)$.

(3) For each of the N words d :

a) Choose a topic $z_n \sim Multinomail(\theta)$.

b) Choose a word w_n from $P(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n .

3.2 Identify Topic Based on KL-Divergence

Kullback-Leibler (KL) divergence is a measure of the difference between the distributions P and Q . Usually P denotes the distribution of observed data, Q denotes a model or an assumed distribution. For the discrete distributions, the KL-divergence between P and Q is defined as Equation (1).

$$D_{KL}(P \parallel Q) = \sum_{i=1}^{n} P_i \log P_i / Q_i \tag{1}$$

We compute the distance between themes, and the distance between theme and Blog sentences using KL-divergence. In the trained LDA model, each theme can be represented by the weights of the feature words. The theme is denoted by $Theme_t = (w_{t1}, w_{t2}, ..., w_{tn})$. With the vector space model, a collection of Blog text sentences can be denoted as $\mathbf{p}_r = (\omega_{r1}, \omega_{r2}, ..., \omega_{rn})$. According to Equation (1), the KL-divergence between $\mathbf{p}_r$ and $Theme_t$ can be computed with Equation (2).

$$D_{KL}(P_r \parallel Theme_t) = \sum_{i=1}^{n} \omega_{ri} \log \omega_{ri} / w_{ti} \tag{2}$$

Because the KL-divergence is non-symmetry, we use Equation (3) to compute the distance in our experiments.

$$D_{KL}(P_r, Theme_t) = (D_{KL}(P_r \parallel Theme_t) + D_{KL}(Theme_t \parallel P_r))/2 \tag{3}$$

4 Sentiment Calculation Based on Hownet Lexicon

The sentiment orientation of Blog themes is based on the sentiment words tagging method. At first some sentiment words with positive and negative polarities are selected as sentiment benchmark words, then the sentiment polarity of other sentiment words are calculated based on the relationship between words in Hownet lexicon, and then the sentiment orientation of a given sentence is determined by judging the sentiment orientation of the individual word or phrase in the sentiment sentence, finally the sentiment orientation of the paragraphs or article are calculated through the sentiment orientation of these sentences.

4.1 Introduction to Hownet Lexicon

Hownet is a concept of Chinese and English words represented as describing objects, which is a general knowledge base to reveal inter-concept relations and relations between the attributes of concepts. Hownet uses a series of sememes to represent every concept. Sememes are organized into a tree structure through contextual relationship. In Hownet the vocabulary similarity is calculated by the similarity between the respective concepts of vocabularies, and the similarity between the concepts is defined by the similarity between sememes. The similarity between sememes can be calculated by the distance between two sememes in the sememe tree.

For two Chinese words W_1 and W_2, if W_1 with n concepts denotes by $\{S_{11}, S_{12}, \ldots, S_{1n}\}$, W_2 with m concepts denotes by $\{S_{21}, S_{22}, \ldots, S_{2m}\}$, the similarity between W_1 and W_2 equal to each word concept similarity of the maximum value, as shown in Equation (4).

$$Sim(W_1, W_2) = \max_{i,j} Sim(S_{1i}, S_{2j}) \qquad i = 1, 2, ..., n \quad j = 1, 2, ..., m \tag{4}$$

The similarity between sememes can be calculated by their path distance p_1 and p_2 in the sememe tree. So the sememe similarity is calculated with Equation (5).

$$Sim(p_1, p_2) = \frac{\alpha}{l + \alpha} \tag{5}$$

Where l is a positive integer which denotes the path length of p_1 and p_2 in sememes tree, α denotes an adjustable parameter.

4.2 Calculation Method for the Sentiment Polarity of Sentiment Word

First, we choose m pairs of words with strong positive and negative polarity as the benchmark words [5], in which the polarity of the positive word is set as +1 and the negative is set -1. The sentiment polarity of each sentiment word is calculated by the similarity between each benchmark word and sentiment word in Hownet. The sentiment polarity of words W is shown as Equation (6). If $O(W) > 0$, that means the sentiment polarity of W is positive. If $O(W) < 0$, that means the sentiment polarity of W is negative, and $O(W) = 0$, that is the sentiment polarity of W is neutral.

$$\mathrm{O}(W) = \sum_{i=1}^{m} \left(Sim(kp_i, W) - Sim(kq_i, W) \right) \tag{6}$$

Where kpi denotes the polarity value of benchmark words with $+1$, kq_i denotes the polarity value of benchmark words with -1, m indicates that there is m pair of benchmark words, $Sim(kp_i, W)$ is equivalent to Equation (4).

We find that if the word is a positive word, the greatest similarity with the word will be get form the benchmark words with $+1$, and if the word is negative word, the greatest similarity with the word will be get from the benchmark words with -1. Based on this idea, we improve Equation (6) to Equation (7).

$$\begin{aligned}\mathrm{O}(W) = \frac{1}{3}\Big(\frac{1}{m}\big(\sum_{i=1}^{m} \left(Sim(kp_i, W) - Sim(kq_i, W) \right)\big) \\ + MAX\,(Sim(kp_j, W)) - MAX\,(Sim(kq_j, W))\Big)\end{aligned} \tag{7}$$

Where $j = 1, 2, ..., m$, $MAX\,(Sim(kp_j, W))$ indicates the maximum similarity value between the sentiment word W with the positive benchmark words, $MAX\,(Sim(kq_j, W))$ indicates the maximum similarity value between the sentiment word W with the negative benchmark words. We use the maximum similarity to improve the sentiment polarity of the sentiment word.

5 Experiments and Evaluation

Because there are not authoritative datasets, the web crawler software MetaSeeker [1] is used to grabs Blog pages from the SINA Blog. We get a dataset with 2000 SINA Blog texts about “real-name system of train tickets (火车票实名制)”. These Blog page content includes the title, label, Blog text and comments. In our experiments, three persons are invited to label the sentiment orientation of the Blog text, which is tagged as positive or negative or neutral. We compare the maximum accuracy value of different methods in our experiments.

5.1 The Sentiment Orientation over the Blog Article Level

We judge the sentiment orientation O(Doc) of the whole Blog text in this experiment at first. O(Doc) is a specific values between -1 and $+1$. In the experiment, we set a

[1] http://www.gooseeker.com/cn/node/product/front

threshold value k. If $-1 \le \mathrm{O(Doc)} < -k$, the blog text is classified to negative, else if $-k \le \mathrm{O(Doc)} \le k$, the blog text is classified to neutral, else if $k < \mathrm{O(Doc)} \le +1$, the blog text is classified to positive. The results of the sentiment classification accuracy with different k are shown in Table 1.

Table 1. Different threshold k sentiment orientation of blog text accuracy

The value of k	0.005	0.01	0.015	0.02	0.025
accuracy（%）	54.33	65.96	73.28	70.03	64.54

From Table 1, we can see that the maximum accuracy is only 73.28%. This result is not enough high. The reason may be that to determine the sentiment orientation of the blog text in document level is “too big”. So we compute the sentiment orientation of multi-aspect blog themes in the next experiments.

5.2 The Sentiment Orientation over Multi-aspect Blog Themes

In this experiment we use LDA model to find Blog themes, and then use KL-divergence to calculate the distance between the Blog themes. The value of M is set as 6. We can roughly summarize the semantic of each theme through analyzing the Blog text. All the themes respectively denote as Theme1 to Theme6. Theme1 can be summed up as “the cost of the real-name system of train tickets(火车实名制投入成本)”, Theme2 is “the real-name system of train tickets will lead to information loss(火车实名制导致信息丢失)”, Theme3 is “the real-name system of train tickets will combat scalpers(火车实名制打击黄牛党)”, Theme4 is “the experiments of the real-name system of train tickets (火车实名制试点)”, and Theme5 is summarized as “the certificates required to buy tickets and get into train station in the real-name system of train tickets(火车实名制进站、买票所需要证件)”. In order to improve the identification accuracy, we also add a Theme6 for other themes which is not included in above five themes.

We used respectively sentence, paragraph and dynamic sliding window to calculate the distance between test text and each theme. The experiment result shows that the accuracy of the first method and the third method are less than 50% , the accuracy of the method for using paragraph can achieve 72.9% . Hence, we use the paragraph as the basic unit to split Blog text in experiment.

To better identify Blog themes based on paragraphs, we set the thresholds ε and θ of D_{KL1} which shown in Equation (8). If both Equation (8) and $D_{KL1} < \theta$ are satisfied, the tested paragraph will be classified into the theme pointed by D_{KL1}.

$$(D_{KL1} - D_{KL2}) / D_{KL1} \ge \varepsilon \tag{8}$$

Where D_{KL2} indicates the distance of the theme which is in the nearest the paragraph, D_{KL1} denotes the distance of the theme which is in the second near the paragraph. The experiment results show that when the value of ε equals to 0.2 and θ equals to 0.05, the accuracy achieves 91.2254%. Table 2 shows the theme identification results. The number of paragraphs which are identified correctly is 1809.

Table 2. Theme identification results in LDA model

Themes	Theme1	Theme2	Theme3	Theme4	Theme5	Theme6
The number of paragraphs with theme labels	213	365	670	408	327	625
The number of Paragraphs identified correctly	201	316	621	399	272	126

With the 1809 paragraphs, we determine their sentiment orientation with Hownet lexicon. All the sentiment orientation of the 1809 paragraphs is labeled beforehand. The number of positive paragraphs is 593, the number of negative paragraphs is 789, and the number of neutral paragraphs is 427. The experiment results show that the sentiment orientation of 1613 paragraphs is identified correctly. The identification accuracy is 89.165%. The identification accuracy of each theme is list in Table 3.

Table 3. Sentiment analysis results of various Theme

Themes	Theme1	Theme2	Theme3	Theme4	Theme5
The number of paragraphs in each theme	201	316	621	399	272
The number of paragraphs identified correctly	179	287	563	326	258
Identification accuracy(%)	89.05	90.82	90.66	81.70	94.85

6 Conclusions

This paper proposes a multi-aspect Blog sentiment analysis based on LDA topic model and Hownet lexicon. This method adopts trained LDA model to split the Blog page text into multi-aspect themes. After that the sentiment word tagging method based on Hownet is used to calculate the sentiment orientation of every Blog theme. So the sentiment orientation of the Blog pages can be represented by the sentiment orientation of multi-aspect Blog themes. The experiment results on SINA Blog dataset show our method can get good topic segments and the sentiment classification performance.

References

1. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424 (2002)
2. Xu, L., Lin, H., Yang, Z.: Text orientation identification based on semantic comprehension. Journal of Chinese Information Processing 21, 96–100 (2007)

3. Du, W., Tan, S., Yun, X., Cheng, X.: A new method to compute semantic orientation. Journal of Compute Research and Development 46, 1713–1720 (2009)
4. Liu, Q., Li, S.: Word similarity computing Based on HowNet. In: The 3th Chinese Lexical Semantic Workshop, CLSW 2002 (2002)
5. Zhu, Y.: Semantic orientation computing based on HowNet. Journal of Chinese Information Processing 20, 14–20 (2006)
6. Whitelaw, C., Garg, N., Argamon, S.: Using appraisal groups for sentiment analysis. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 625–631. ACM, Bremen (2005)
7. Wiebe, J., Lin, W.H., Wilson, T., Hauptmann, A.: Which Side are You on?:identifying perspectives at the document and sentence levels. In: Proceedings of the Tenth Conference on Natural Language Learning, pp. 109–116. ACM, New York City (2006)
8. Tang, H., Tan, S., Cheng, X.: Research on sentiment classification of Chinese reviews based on supervised machine learning techniques. Journal of Chinese Information Processing 21, 88–94 (2007)
9. Liao, X., Cao, D., Fang, B., Xu, H., et al.: Research on Blog opinion retrieval based on probabilistic inference model. Journal of Compute Research and Development 46, 1530–1536 (2009)
10. Xu, J., Ding, Y., Wang, X.: Sentiment classification for Chinese news using machine learning methods. Journal Chinese Information Processing 21, 95–100 (2007)
11. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)

Redundant Feature Elimination by Using Approximate Markov Blanket Based on Discriminative Contribution

Xue-Qiang Zeng[1], Su-Fen Chen[2], and Hua-Xing Zou[1]

[1] Computer Center, Nanchang University, Nanchang 330031, China
{xqzeng,hxzou}@ncu.edu.cn

[2] Department of Computer Science and Technology, Nanchang Institute of Technology,
Nanchang 330099, China
sufenchen@foxmail.com

Abstract. As a high dimensional problem, it is a hard task to analyze the text data sets, where many weakly relevant but redundant features hurt generalization performance of classifiers. There are previous works to handle this problem by using pair-wise feature similarities, which do not consider discriminative contribution of each feature by utilizing the label information. Here we define an Approximate Markov Blanket (AMB) based on the metric of DIScriminative Contribution (DISC) to eliminate redundant features and propose the AMB-DISC algorithm. Experimental results on the data set of Reuter-21578 show AMB-DISC is much better than the previous state-of-arts feature selection algorithms considering feature redundancy in terms of Micro$_{avg}$F1 and Macro$_{avg}$F1.

1 Introduction

Feature selection, a process of choosing a subset of features from the original ones, is frequently used as a preprocessing technique in text mining. It has been proved effective in reducing dimensionality, improving mining efficiency, increasing mining accuracy, and enhancing result comprehensibility [1–3]. However, redundant features can not be completely eliminated by commonly used feature selection algorithms. For instance, as to analysis of text data sets, whose speciality is the huge amount of words (features), it is believed that there are many weakly relevant but redundant words among the full word set. Preserving the most discriminative words and reducing other irrelevant and redundant ones is the target of feature selection. However, because the interactions and correlations among features are not considered, common feature selection algorithms fail to remove redundant words.

The issue of redundancy among features is recently raised in the literatures of feature selection [4, 5]. Researchers have proposed several algorithms to reduce the redundancy among features. Ding and Peng proposed the minimum Redundancy-Maximum Relevance (mRMR) algorithm [5], which requires that selected discriminative features are maximally dissimilar to each other. Yu and Liu proposed the Fast Correlation-Based Filter (FCBF) algorithm [4], which eliminates redundant features by iterative selecting predominant features from relevant ones.

However, all these algorithms estimated the pair-wise redundancy among features by using normal feature distance (similarity) measures, which could not work properly. Because without consideration of the label information, the pair-wise redundancy

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 139–146, 2011.

scores solely calculated by the given two features do not faithfully reflect the similarity of their discriminative abilities. For example, two highly correlated features, whose differences are minor but happen to own different critical discriminative ability, may be considered as a pair of redundant features by the previous commonly used metrics. Hence, reducing any one of them will decrease the final classification accuracy. Zeng *el al.* proposed a novel metric of redundancy based on DIScriminative Contribution (DISC) which estimates the feature similarity by explicitly building a linear classifier on each feature [6]. Based on DISC, we define an Approximate Markov blanket (AMB) for the searching of redundant features, and propose an algorithm named AMB-DISC, which produces a compact feature set with high performance. We compare AMB-DISC with other redundant feature selection algorithms on the data set of Reuter-21578, and demonstrate its the outstanding performance.

This paper is organized as follows. Section 2 describes the metric of discriminative contribution. In Section 3, our proposed algorithm is presented in detail. Data sets and experiment settings are described in Section 4. We show the results and discussions in Section 5. Finally, conclusions are given in Section 6.

2 The Metric of Discriminative Contribution

Notions of feature redundancy are normally in terms of feature correlation. It is widely accepted that two features are redundant to each other if their values are completely correlated. But in fact, it may not be so straightforward to determine feature redundancy when a feature is correlated with a set of features. A widely used way is approximate the redundancy of feature set by only considering the pair-wise feature redundancy, *i.e.* the algorithms we mentioned above. The success of these algorithms relies much on the effectiveness of pair-wise similarity measures. However, exist pair-wise similarity measures do not faithfully reflect the similarity of their discriminative abilities. For example, two highly correlated features, whose differences are minor but happen to causing different critical discriminative power, maybe be considered as a pair of redundancy features. Reducing any one of them will decrease the classification accuracy.

One possible way of identifying the redundancy between feature's predictive ablilties is comparing the distributions of discriminative powers between them. Zeng *et al.* proposed a metric of redundancy based on DIScriminative Contribution (DISC)[6]. After building classifier on each sole feature, DISC refeeding the whole training samples back to estimate the performance of each classifier. A high self-training accuracy score means the corresponding feature has great discriminative power. In most cases, only a part of the training samples can be correctly separated by this classifier, because only one feature is used. Based on these classifiers, the difference of classification distribution is used to estimate the discriminative contribution. Given two features x and y, two classifiers $classifer_x$ and $classifer_y$ are constructed. The differences of the correctly classified training samples of $classifer_x$ and $classifer_y$ are recorded in Table 1.

In order to eliminate redundant features, we want to examine whether the contribution of the additional feature is significant to the given one. The additional feature is considered as redundant one only when its contribution is tiny. Based on the pair-wise

Table 1. Discriminative Cross Table

$classifer_y$ $classifer_x$	true	false
true	a	b
false	c	d

discriminative contribution, the DIScriminative Contribution (DISC) value of x to y, which represents the y's redundancy to x, is defined as follows.

$$\mathrm{DISC}(x, y) = \frac{d}{c + d} \tag{1}$$

where $c + d$ is the samples which could not be discriminated by $classifer_x$, d is the samples which could not be rightly classified even by the collaboration of $classifer_x$ and $classifer_y$. So the proportion of $\frac{d}{c+d}$ is the useless extent of $classifer_y$ to $classifer_x$.

The DISC score varies from 0 to 1, and takes 1 only when c is 0. The value of c takes 0 when all the discriminative power of $classifer_y$ is coved by $classifer_x$ and $classifer_y$ has no discriminative contribution to $classifer_x$. In this case, y is considered completely redundant to x. On the other hand, when the DISC value is 0, both a and d are 0, the discriminative power of y is supposed complementary to that of x. The computation complexity of DISC is $O(n)$.

3 The AMB-DISC Algorithm

It is not so straightforward to determine feature redundancy when a feature is correlated with a set of features. We need formally define feature redundancy in order to devise an approach to explicitly identify and eliminate redundant features. We first introduce the definition of a feature's Markov blanket given by Koller and Sahami [7].

Definition 1. *Given a feature set F, a label C, and a feature $F_i \in F$, let $M_i \in F(F_i \notin M_i)$, M_i is said to be a Markov blanket for F_i iff*

$$P(F - M_i - F_i, C|F_i, M_i) = P(F - M_i - F_i, C|M_i)$$

The Markov blanket condition requires that M_i subsume not only the information that F_i has about C, but also about all of the other features. It is pointed out by Koller and Sahami [7] that the optimal subset is obtained by a backward elimination procedure, known as Markov blanket filtering: let G be the current set of features ($G = F$ in the beginning), at any phase, if there exists a Markov blanket for F_i within the current G, F_i is removed from G. It is proved that this process guarantees a feature removed in an earlier phase will still find a Markov blanket in any later phase, that is, removing a feature in a later phase will not render the previously removed features necessary to be included in the optimal subset. Then definition of redundant feature is given:

Definition 2. *Let G be the current set of features, a feature is redundant and hence should be removed from G iff it is weakly relevant and has a Markov blanket M_i within G.*

The definition of redundant features relies on the computation of Markov blanket, which is hard to be solved in real problem. In practical, when it comes to approximately determine feature redundancy, the key is to find approximate Markov blankets for the selected relevant features. Similar as what has been done in the FCBF algorithm [4], we define a novel approximate Markov blanket based on the discriminative contribution of feature. The definition is given as follows

Definition 3. *For two features F_i and F_j $(i \neq j)$, F_j forms an approximate Markov balnket for F_i iff $CHI_j \geq CHI_i$ and $DISC(j,i) \geq \delta$*

In definition 3, CHI_i is the correlation between any feature F_i and the class C, and it is believed that a feature with a larger CHI value contains more information about the class than a feature with a smaller CHI value. The definition of CHI_i is given as follows.

$$CHI_i = \frac{n \times (AD - BC)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}$$

where A and B are the non-zero value counts of F_i for positive and negative examples respectively, C and D are the zero value counts for positive and negative examples respectively, and n is the size of training set.

We determine the existence of an approximate Markov blanket between a pair of features F_i and F_j based on their DISC value. When $DISC(j,i) \geq \delta$, we believe F_i's discriminative contribution for F_j is tiny. In other words, feature F_j can form an approximate Markov blanket for feature F_i. δ is a threshold given by users.

The approximation algorithm for redundancy analysis presented above is realized by an algorithm, named AMB-DISC (Approximate Markov Blanket based on DISC), which is illustrated in Figure 1. Firstly, the features are ordered by their CHI scores. As we usually want to retain the more discriminative one between two redundant features, AMB-DISC tries to preserve the top CHI score ranked features. Then, AMB-DISC uses two nested iterations to eliminate redundant features whose discriminative abilities are approximate Markov blanket covered by a higher ranked features. As the inner iteration works in a backward way, it is guaranteed that each removed feature can find an approximate Markov blanket cover in the final selected feature set.

The computational complexity of AMB-DISC is $O(npk)$, where k is number of selected features. When k is much smalller than p, the computational complexity of AMB-DISC is $O(np)$, which is the same as that of commonly used filter algorithms. As we konw, in many large data sets, discriminative features with high CHI scores are always rarely compared with other ones. Most redundant or weakly redundant features are easily to be approximate Markov blanket covered by few high discriminative features. So, the size of final compact feature set produced by AMB-DISC is much smaller than that of original feature set in most cases, which are examined in our experiments later.

4 Experimental Settings

The Reuters-21578 collection is used in our experiments, which is divided into a training set and a test set by the ModApte split as in the previous studies [8]. By removing

Algorithm 1. The AMB-DISC Algorithm

```
Input:  Feature set $X = [x_1, x_2, \ldots, x_p]$
        Target label $C$
        Threshold $\delta$
Output: Selected Feature subset $S$

begin
$S' \Leftarrow X$
for $i = 1$ to $p$ do
    calculate $CHI_i$ for $X_i$
end for
order $S'$ in descending $CHI_i$ value
for $j = 1$ to the size of $S'$ do
    for $i$ =the size of $S'$ to $j + 1$ do
        if DISC($S'_j, S'_i$) > $\delta$ then
            remove feature $i$ from $S'$
        end if
    end for
end for
$S \Leftarrow S'$
end
```

some corrupted documents, we obtain 7,770 training documents and 3,019 test documents, which is the same with the data preprocess in some related works [8, 9]. There are 90 different categories. We preprocess the data in a formal way: all numbers and stopwords are removed, words are converted into lowercase, word stemming is performed using the Porter stemmer, some noisy words are removed. In the end, we obtain 6,883 unique terms.

The multi-label classification problem here is solved by training 90 binary classifiers for each class. The feature selection model is built and applied for each classifier separated. To make the conclusions sound, five widely used classification algorithms are employed in our experiments, which are list as follows.

- **5NN** k nearest neighbor with $k = 5$.
- **C45** The C4.5 decision tree.
- **LG** Logistic regression.
- **SMO** Support vector machines using Sequential Minimal Optimization.
- **SVM** Linear support vector machine with $c = 10$.

For the evaluation, we use the standard Macro$_{avg}$F1 and Micro$_{avg}$F1 measure [8]. Macro$_{avg}$F1 measure gives the same weight to all categories, and thus it is equally influenced by the performance of rare categories. On the contrary, Micro$_{avg}$F1 measure is dominated by the performance of common categories. All the experiments are performed on a PC machine with P4 2.0G CPU and 2G RAM. The programming language is JAVA, the open source machine learning project of WEKA [10] is used.

5 Results and Discussions

We compare our algorithm with several typical feature selection algorithms, which are list as follows.

- **CHI** Commonly used filter feature selection algorithm with CHI values. Some top scored features are retained for each separate classifer. The number of selected feature has been tried for 10, 20, 50, 100, 150, 200, 300, 400 and 500. We found 500 is the best one, which is also used to compare with other algorithms.
- **CFS [11]** The Correlation-based Feature Selection (CFS) algorithm exploits best-first search based on some correlation measure which evaluates the goodness of a subset. The standard linear correlation is used in our experiments.
- **mRMR [5]** The minimum Redundancy-Maximum Relevance (mRMR) algorithm requires that selected discriminative features are maximally dissimilar to each other. The maximal number of selected feature is a parameter, which has been tried for 10, 20, 50, 80 and 100 in our experiments. The best result of mRMR is got when 80 features are selected.
- **FCBF [4]** The Fast Correlation-Based Filter (FCBF) algorithm, which eliminates redundant features by iterative selecting predominant features from relevant ones.
- **AMB-DISC** Our proposed algorithm eliminates redundant features by approximate Markov blanket and discriminative contribution between features. The parameter of δ is set to 0.95.

Comparative results of the size of the selected feautre set is given in Tables 2, where Dim.±std is the statistical mean feature number with its standard deviation over 90 classes. The comparative Micro$_{avg}$F1 and Macro$_{avg}$F1 results are showed in Tables 3 and Tables 4 respectively. The last row is the average scores over five different classifiers.

Table 2. The number of selected features by using different algorithms

	CHI	CFS	mRMR	FCBF	AMB-DISC
Dim.	500.00	18.11±9.22	61.69±30.39	62.63±138.29	85.90±23.57

It can be seen from Table 2 that the numbers of selected features by using CFS, mRMR, FCBF and AMB-DISC are both much smaller than that of CHI, which indicating redundant features can not be excluded by filter algorithms without considering redundancy. It is necessary to do redundancy analysis to get an optimal or sub-optimal compact feature set.

From Table 3~4, we can see that no feature selection algorithm has overwhelming performance with all classifiers. In average, AMB-DISC is the best one. Especially for the Micro$_{avg}$F1 scores, the averaged score of AMB-DISC is much better than that of all other algorithms. Furthermore, AMB-DISC always has the top performance with all classifiers. For the averaged Macro$_{avg}$F1 scores, AMB-DISC is also the winner. But the improvement is not so significant, and AMB-DISC is worse than CHI in some cases

Table 3. Comparative results of Micro$_{avg}$F1 by using different algorithms

	CHI	CFS	mRMR	FCBF	AMB-DISC
5NN	0.683	0.798	0.798	0.774	0.838
C45	0.807	0.799	0.806	0.784	0.807
LG	0.838	0.814	0.827	0.779	0.865
SMO	0.861	0.812	0.822	0.780	0.867
SVM	0.854	0.825	0.836	0.795	0.866
AVG.	0.808	0.809	0.818	0.782	0.849

Table 4. Comparative results of Macro$_{avg}$F1 by using different algorithms

	CHI	CFS	mRMR	FCBF	AMB-DISC
5NN	0.285	0.376	0.356	0.342	0.516
C45	0.452	0.436	0.446	0.405	0.422
LG	0.452	0.451	0.401	0.425	0.543
SMO	0.555	0.364	0.352	0.360	0.488
SVM	0.590	0.501	0.471	0.452	0.603
AVG.	0.467	0.425	0.405	0.397	0.514

i.e. with the classifiers of 5NN and SMO. This shows that removing redundant features by discriminative contribution has obvious postive effects on classification. The enhancements are more significant for common classes than rare classes. We believe this is because identifing redundant features needs adequate label information, and feature selection is easy to be trapped by noises when positive examples are rare.

The previous state-of-arts feature algorithms considering redundancy, including CFS, mRMR and FCBF, are both not better than the filter algorithm of CHI. For the averaged Micro$_{avg}$F1 scores, the performances of CFS, mRMR and FCBF are similar with that of CHI, although the corresponding feature numbers are much less. But for the Macro$_{avg}$F1 scores, CFS, mRMR and FCBF are obviously worse than CHI, which representing some useful features are mistakenly treated as redundant ones for rare classes. This also confirms that identifing redundant features needs adequate label information. By inadequate redundany meatures, removing redundant features obviously decreases the classification performance. When classification is the final task, we believe redundancy among features should be measured by the difference among discriminative powers of features not the numerical values.

6 Conclusions

Redundant feature elimination is an important topic in the field of feature selection for text mining. However, the measurement of feature redundancy is still an open problem. Existing feature selection algorithms usually employ pair-wise feature similarity to represent feature redundancy. However, state-of-arts similarity measures including linear and non-linear ones calculate the redundancy only by the feature's numerical values without considering the label information, which is vital to the classification, while not used to estimate the discriminative contribution of the differences of features. Here we define an approximate Markov blanket (AMB) based on the metric of DIScriminative Contribution (DISC) for the searching of redundant features, and propose an algorithm AMB-DISC. Experimental results on the data set of Reuter-21578 show AMB-DISC produces much more compact feature set than commonly used filter algorithms, and show obviously better performance than other state-of-arts feature selection algorithms considering redundancy, *i.e.* CFS, mRMR and FCBF.

Acknowledgment. This work was supported by the Natural Science Foundation of Jiangxi Provincial, China (Grant No. 2010GQS0159).

References

1. Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence 97(1-2), 245–271 (1997)
2. Liu, H., Dougherty, E., Dy, J., Torkkola, K., Tuv, E., Peng, H., Ding, C., Long, F., Berens, M., Parsons, L., Zhao, Z., Yu, L., Forman, G.: Evolving feature selection. IEEE Intelligent Systems 20(6), 64–76 (2005)
3. Zhu, S., Wang, D., Yu, K., Li, T., Gong, Y.: Feature selection for gene expression using model-based entropy. IEEE Transactions on Computational Biology and Bioinformatics 7(1), 25–36 (2010)
4. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)
5. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1226–1238 (2005)
6. Zeng, X.Q., Li, G.Z., Yang, J.Y., Yang, M.Q., Wu, G.F.: Dimension reduction with redundant genes elimination for tumor classification. BMC Bioinformatics 9(suppl 6), S8 (2008)
7. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284–292 (1996)
8. Yang, Y., Liu, X.: A re-examination of text categorization methods. In: SIGIR 1999: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42–49. ACM Press, New York (1999)
9. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Survey 34(1), 1–47 (2002)
10. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
11. Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In: International Conference on Machine Learning, pp. 359–366 (2000)

Synchronization of Hyperchaotic Rossler System and Hyperchaotic Lorenz System with Different Structure

Yi-qiang Wei and Nan Jiang

Department of mathematics, Taiyuan University of Technology,
030024 Taiyuan, China
wei_yiqiang@163.com, exiu5815@sina.com

Abstract. This paper studies synchronization between hyperchaotic Rossler system and hyperchaotic Lorenz system with unknown parameters. Based on Lyapunov stability theory, active synchronization and adaptive synchronization make the different systems achieve synchronization. And numerical simulations show the effectiveness and feasibility of these methods.

Keywords: hyperchaotic system, active synchronization, adaptive synchronization, Lyapunov stability theory component.

1 Introduction

In the last few years, chaos synchronization has received a lot of attention among scientists from variety of research field. The first idea of synchronizing two identical chaotic systems with different initial conditions was introduced by Pecora and Carrols[1]. Recently, many control methods have been developed to achieve chaos synchronization in two chaotic systems, such as adaptive control[2-4], linear balanced feedback control[5], impulsive control[6, 7], sliding mode control[8], fuzzy control[9], backstepping control[10], and so on. The aforementioned methods and some other existing synchronization methods mainly concerned the synchronization of two identical chaotic systems with known parameters or unknown parameters. However, it is hardly the case that a lot of hyperchaotic systems exist in real world, moreover, the system's parameters are not exactly known in priori. Therefore, synchronization between two different unknown hyperchaotic systems is more useful in real-life application.

In this paper, we propose a simple and general approach to synchronization hyperchaotic Rossler system and hyperchaotic Lorenz system with unknown parameters. Based on Lyapunov stability theory, active synchronization and adaptive synchronization make the different systems acheive synchronization. Simulation results show that the proposed method can be successfully used in synchronization of chaotic systems.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 147–153, 2011.

2 Active Synchronization between Hyperchaotic Rossler System and Hyperchaotic Lorenz System

For acheive synchronization between hyperchaotic Rossler system and hyperchaotic Lorenz system, we consider hyperchaotic Rossler system as the drive system

$$\begin{cases} \dot{x}_1 = -x_2 - x_3 \\ \dot{x}_2 = x_1 + ax_2 + x_4 \\ \dot{x}_3 = b + x_1 x_3 \\ \dot{x}_4 = -cx_3 + dx_4 \end{cases} \tag{1}$$

and hyperchaotic Lorenz system as the response system

$$\begin{cases} \dot{y}_1 = a_1(y_2 - y_1) + u_1 \\ \dot{y}_2 = b_1 y_1 + c_1 y_2 - y_1 y_3 + y_4 + u_2 \\ \dot{y}_3 = -d_1 y_3 + y_1 y_2 + u_3 \\ \dot{y}_4 = -r_1 y_1 + u_4 \end{cases} \tag{2}$$

where $u_1, u_2 . u_3, u_4$ are controllers. We get the error system

$$\begin{cases} \dot{e}_1 = a_1 y_2 - a_1 y_1 + x_2 + x_3 + u_1 \\ \dot{e}_2 = b_1 y_1 + c_1 y_2 - y_1 y_3 + y_4 - x_1 - ax_2 - x_4 + u_2 \\ \dot{e}_3 = -d_1 y_3 + y_1 y_2 - b - x_1 x_3 + u_3 \\ \dot{e}_4 = -r_1 y_1 + cx_3 - dx_4 + u_3 \end{cases} \tag{3}$$

where $e_i = y_i - x_i$. The controllers were designed for synchronization drive system and response system, it is designed as

$$\begin{cases} u_1 = a_1 x_1 - (a_1 + 1)x_2 - x_3 + V_1 \\ u_2 = y_1 e_3 - e_4 - (b_1 - 1)x_1 - (c_1 - a)x_2 + y_1 x_3 + V_2 \\ u_3 = -y_1 e_2 + (d + x_1)x_3 - y_1 x_2 + b + V_3 \\ u_4 = r_1 x_1 - cx_3 + dx_4 + V_4 \end{cases} \tag{4}$$

where V_1, V_2, V_3 are the control input based on functions e_1, e_2, e_3. Substituting Eq. (4) into Eq. (3), then we have a new error system

$$\begin{cases} \dot{e}_1 = -a_1 e_1 + a_1 e_2 + V_1 \\ \dot{e}_2 = b_1 e_1 + c_1 e_2 + e_4 + V_2 \\ \dot{e}_3 = -d e_3 + V_3 \\ \dot{e}_4 = -r_1 e_1 + V_4 \end{cases} \tag{5}$$

So the error system(3) is replaced by a linear syetem(5), where V_1, V_2, V_3 are the control input based on functions e_1, e_2, e_3. The drive syetem(1) and response system(2) are said to be synchronized, if the trivial solution of error system trajectory is asymptotically stable in mean square, in the sense that

$$\lim_{t\to\infty} e_i(t) = \lim_{t\to\infty} \left| y_i(t) - x_i(t) \right| = 0$$

For V_1, V_2, V_3 have some choose, thus, we may denote

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = A \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix}$$

where A is a constant matrix. For the error system(5) is asymptotically stable in mean square, we select A 's elements such as all eigenvalue of system(5) has negative part. For convenience, we choose

$$A = \begin{bmatrix} a_1 - 1 & 0 & 0 & 0 \\ -b_1 & -c_1 - 1 & 0 & 1 \\ 0 & 0 & d - 1 & 0 \\ r_1 & 0 & 0 & -1 \end{bmatrix}$$

In this case, the error system(5)'s eigenvalue is $-1,-1,-1,-1$, the error e_i converges to 0 as time t approaches infinity. Therefore, hyperchaotic Rossler system and hyperchaotic Lorenz system acheive active synchronization.

3 Adaptive Synchronization between Hyperchaotic Rossler System and Hyperchaotic Lorenz System

3.1 The First Method

In this section, in oder to realize the synchronization between the drive system(1) and the response system(2), we will choose a general and suitable synchronization controller, and then develop theoretical results of the adaptive synchron-ization scheme.

Theorem 1. If the nonlinear controller u is designed as

$$\begin{cases} u_1 = a_1 x_1 - (a_1 + 1)x_2 - x_3 - b_1 e_2 + r_1 e_4 \\ u_2 = y_1 e_3 - (b_1 - 1)x_1 - (c_1 - a)x_2 + y_1 x_3 - a_1 e_1 \\ u_3 = -y_1 e_2 + (d + x_1)x_3 - y_1 x_2 + b \\ u_4 = r_1 x_1 - cx_3 + dx_4 + ke_4 - e_2 \end{cases} \tag{6}$$

the adaptive laws of parameters are taken as $\beta = ke_4$ (where k is a constant), then the response system(2) can synchronize the drive system(1) asymptotically, where parameters are all unknown.

Proof. Substituting Eq. (6) into Eq. (3), then we get the error system

$$\begin{cases} \dot{e}_1 = -a_1 e_1 + a_1 e_2 - b_1 e_2 + r_1 e_4 \\ \dot{e}_2 = b_1 e_1 + c_1 e_2 + e_4 - a_1 e_1 \\ \dot{e}_3 = -de_3 \\ \dot{e}_4 = -r_1 e_1 + ke_4 - e_2 \end{cases}$$

Define a Lyapunov function as

$$V(t) = \frac{1}{2}(e_1^2 + e_2^2 + e_3^2 + e_4^2) \tag{7}$$

Differentiating both sides of Eq. (7) yields:

$$\begin{aligned} \dot{V}(t) = {} & e_1(-a_1 e_1 + a_1 e_2 - b_1 e_2 + r_1 e_4) + \\ & e_2(b_1 e_1 + c_1 e_2 + e_4 - a_1 e_1) + e_3(-de_3) + e_4(-r_1 e_1 + ke_4 - e_2) \\ = {} & -a_1 e_1^2 + c_1 e_2^2 - de_3^2 + ke_4^2 < 0 \end{aligned}$$

Here it is assumed that $a_1 > 0, c_1 < 0$, $d > 0, k < 0$, then differential quotient $\dot{V}(t) \le 0$. Based on Lyapunov stability theory, the error trajectories converge to zero. So under the nonlinear controller u and adaptive laws of parameters, the response system (2) can synchronize the drive system (1).

3.2 The Second Method

Theorem 2. If the nonlinear controller u is designed as

$$\begin{cases} u_1 = -\hat{a}_1 y_2 + \hat{a}_1 y_1 - x_2 - x_3 - k_1 e_1 \\ u_2 = -\hat{b}_1 y_1 - \hat{c}_1 y_2 + y_1 y_3 - y_4 + x_1 + \hat{a} x_2 + x_4 - k_2 e_2 \\ u_3 = \hat{d}_1 y_3 - y_1 y_2 + b + x_1 x_3 - k_3 e_3 \\ u_4 = \hat{r}_1 y_1 - \hat{c} x_3 + \hat{d} x_4 - k_4 e_4 \end{cases} \tag{8}$$

the adaptive laws of parameters are taken as

$$\begin{aligned} \dot{\hat{a}}_1 &= e_1(y_2 - y_1) & \qquad \dot{\hat{r}}_1 &= -y_1 e_4 \\ \dot{\hat{b}}_1 &= y_1 e_2 & \dot{\hat{a}} &= -x_2 e_2 \\ \dot{\hat{c}}_1 &= y_2 e_2 & \dot{\hat{c}} &= x_3 e_4 \\ \dot{\hat{d}}_1 &= -y_3 e_3 & \dot{\hat{d}} &= -x_4 e_4 \end{aligned}$$

then the response system(2) can synchronize the drive system (1) asymptotically, where parameters k_1, k_2, k_3, k_4 are positive and unknown, and $\hat{a}_1, \hat{b}_1, \hat{c}_1, \cdots, \hat{c}, \hat{d}$ is the estimator of $a_1, b_1, c_1, \cdots, c, d$.

Proof. Substituting Eq. (8) into Eq. (3), then we get the error system

$$\begin{cases} \dot{e}_1 = \tilde{a}_1 y_2 - \tilde{a}_1 y_1 - k_1 e_1 \\ \dot{e}_2 = \tilde{b}_1 y_1 + \tilde{c}_1 y_2 - \tilde{a} x_2 - k_2 e_2 \\ \dot{e}_3 = -\tilde{d}_1 y_3 - k_3 e_3 \\ \dot{e}_4 = -\tilde{r}_1 y_1 + \tilde{c} x_3 - \tilde{d} x_4 - k_4 e_4 \end{cases}$$

let $\tilde{a}_1 = a_1 - \hat{a}_1$, $\tilde{b}_1 = b_1 - \hat{b}_1$, $\tilde{c}_1 = c_1 - \hat{c}_1$, $\cdots \tilde{c} = c - \hat{c}$, $\tilde{d} = d - \hat{d}$. Construct the following Lyapunov function

$$V = \frac{1}{2}(e_1^2 + e_2^2 + e_3^2 + e_4^2 + \tilde{a}_1{}^2 + \tilde{b}_1{}^2 + \tilde{c}_1{}^2 + \tilde{d}_1{}^2 + \tilde{r}_1^2 + \tilde{a}^2 + \tilde{c}^2 + \tilde{d}^2)$$

The stochastic derivative of V along trajectories of (9) can be obtained as follows

$$\begin{aligned} \dot{\hat{V}} &= e_1\dot{e}_1 + e_2\dot{e}_2 + e_3\dot{e}_3 + e_4\dot{e}_4 + \tilde{a}_1\dot{\tilde{a}}_1 + \tilde{b}_1\dot{\tilde{b}}_1 + \tilde{c}_1\dot{\tilde{c}}_1 + \tilde{d}_1\dot{\tilde{d}}_1 + \tilde{r}_1\dot{\tilde{r}}_1 + \tilde{a}\dot{\tilde{a}} + \tilde{c}\dot{\tilde{c}} + \tilde{d}\dot{\tilde{d}} \\ &= -k_1 e_1^2 - k_2 e_2^2 - k_3 e_3^2 - k_4 e_4^2 \end{aligned}$$

So differential quotient $\dot{V}(t) \le 0$, the response system(2) can synchronize the drive system(1).

Equations should be punctuated in the same way as ordinary text but with a small space before the end punctuation mark.

4 Numerical Simulation

In this section, numerical simulations show the effectIve-ness and feasibility of these methods. The hyperchaotic Rossler system and hyperchaotic Lorenz system take the parameters $a = 0.25$, $b = 3$, $c = 0.5$, $d = 0.05$, $a_1 = 35$, $b_1 = 7$, $c_1 = 12$, $d_1 = 3$, $r_1 = 5$. When t=10, $e(0) = [-5,4,-7,2,]$, $e(0) = [-8,4,1,2]$, $e(0) = [2,0,1,-3]$, it is observed that the error trajectories converge to zero.

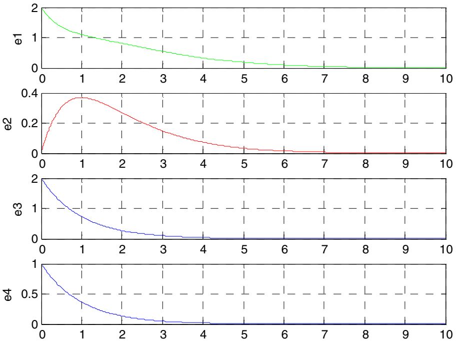


Fig. 1. The synchronization errors with active method

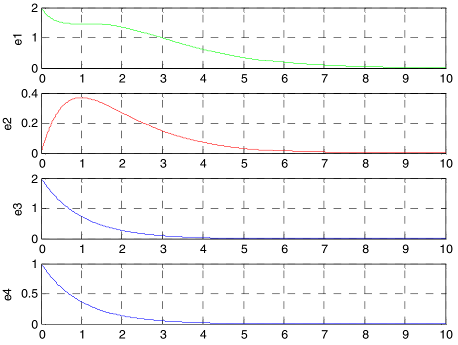


Fig. 2. The synchronization errors with adaptive method (1)

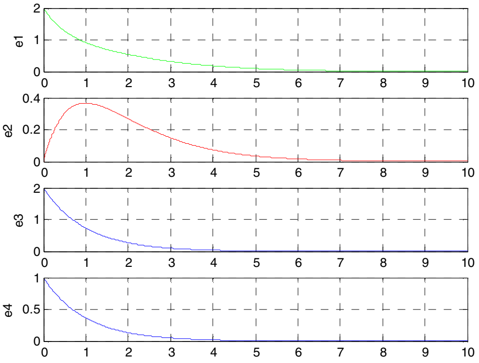


Fig. 3. The synchronization errors with adaptive method (2)

References

1. Carroll, T. L., Pecora, L.M.: Synchronization in chaotic systems. Phys. Rev. Lett. 64, 821–824 (1990)
2. Ge, Z.-M., Lee, J.-K.: Chaos synchronization and parameter identification for gyroscope system. Appl. Math. Comput. 163(2), 667–682 (2005)
3. Efimov, D.-V.: Dynamical adaptive synchronization. Int. J. Adaptive Control Signal Process 20, 491–507 (2006)
4. Hu, J., Chen, S.: Adaptive control for anti- synchronization of Chua's chaotic system. Phys Lett. A 339(6), 455–460 (2005)
5. Chen, H.-H.: Global synchronization of chaotic systems via linear balanced feedback control. Appl. Math. Comput. 186(1), 923–931 (2007)
6. Chen, S.-H., Yang, Q., Wang, C.-P.: Impulsive control and synchronization of unified chaotic system. Chaos Soliton Fract. 20(4), 751–758 (2004)
7. Yan, J.-J., Lin, J.-S., Liao, T.-L.: Synchronization of a modified Chua's circuit system via adaptive sliding mode control. Chaos Soliton Fract. 36(1), 45–52 (2008)
8. Yau, H.-T., Kuo, C.-L., Yan, J.-J.: Fuzzy sliding mode control for a class of chaos. Int. J. Nonlin. Sci. Numer simulate 17(3), 333–338 (2006)
9. Wang, C., Ge, S.-S.: Adaptive synchronization of uncertain chaotic systems via backstepping design. Chaos Soliton Fract. 12(7), 199–206 (2001)
10. Wang, C., Ge, S.-S.: Synchronization of two uncertain chaotic systems via backstepping. Int. J. Bifurcat. Chaos 11, 1743–1751 (2001)

Research of Matrix Clustering Algorithm Based on Web User Access Pattern

Jian Bao

Application Technology College, Liaoning Technical University, Fuxin, China
lntubj@gmail.com

Abstract. It is of great significance that summarizing the regular pattern of the user along the URL to find and browse the Web, mining user browsing patterns to help users reach the target page quickly for realizing the personalized navigation of search engine. In order to provide personalized service, an optimized matrix clustering algorithm is proposed, which can cluster the page users access, analysis and study the laws in the Web log records to improve performance and organizational structure of Web site according to browsing patterns of user accessing to Web, understand the user behavior, find user browsing patterns. The Experiment results shows that the algorithm has good practicability with accurately reflecting the Web visits.

Keywords: User Access Pattern, Web Mining, User Access Matrix, Matrix Clustering.

1 Introduction

Web contains vast amounts of information, but its structure is complex, which is difficult for user to obtain information of interest, while browsing speed is also difficult to guarantee. It become an urgent and important subject that how to analysis requirements of user effectively and help users find information of interest resources. For this purpose, it should make the site according to user behavior patterns optimize organizational structures and forms to improve the quality of Web service, which will greatly facilitate users.

Web log contains the information of user's access, although different users at different times may have different access modes, its long-term trend should be stable, and in another word the interest should be reflected in the behavior of long-term access [1]. In a certain time interval, the behavior of user's access can be recorded by a set of pages the user clicks when accessing the Web site, which are the composition of user session. All the user session in this period constitutes the access path set, and studying the similarity of between the access path set of user session will find user access patterns to the access behavior of user groups. Information such as some about user groups under similar sites will be found by user session clustering [2], which shows the law of behavior that user groups access Web, provide a reference site for optimizing organizational structure, and reduce user access time. It also can provide targeted resources to achieve personalized service.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 154–159, 2011.

2 Web User Access Patterns

User access patterns are the access path which consist of the hyperlink that users click on when the user access site [3]. If different users have the same access sequence of pages hyperlink it means user access behavior has a certain similarity, which is an abstract to user access and it can be viewed as pages of knowledge. Path clustering is to find access behavior of users to access the path shown by the class of set process. Each collection represents one class user access patterns which has similar path. Clustering them can obtain the statistical sense of the dynamic user access patterns and needs, and then to adjust the site structure and provide the basis for further content organization.

Web access pattern mining is the use of data by processing the Web to find user access patterns and understand user behavior. User access patterns mining process is through the use of data mining from the Web access patterns in the process of automatic extraction [4]. The preferences of user access patterns is not only reflected in its sequence of browsing the Web on the path, but also reflected in the user access in Web page timing. Cluster analysis of user groups with similar characteristics help understand the user access patterns [5]. Therefore, by mining the information of user access in a period, similar user groups and relevant page information in this site will be found.

3 Algorithm Description

3.1 Web User Access Patterns Mining Algorithm

Improved matrix clustering algorithm is applied in this paper, which is relatively simple and extensive. Less demanding of its log file, no requirement of its session identification and transaction identification and no complex data make it more convenient.

After preprocessing the original Log data, with L=<*ip*, *uid*, *url*, *time*> is expressed in Web server logs. Which, *UserID* that Web users *ID*, *IP* represents the users *IP* address, *url* that the users requested *url*, *time* indicates that the corresponding browsing time. Then, it's further processing reflect the user browsing behavior within a certain period of time [6].

Definition 1: The user browsing behavior *A*: records of the information user browsing the site left behind, it has shown in Equation 1. *2 (n +1)* tuple:

$$A =< UseID_A, IP_A, \{(l_A, url, hits)\}^n > \tag{1}$$

Among, $l_A\_useid=UserID_A$, $l_A \in L$, $l_A\_ip=IP_A$, $n \geqslant 1$, *hits* represent the times user $UserID_A$ browsing the page l_A_useid.

Definition 2: Website model G: Website can be viewed as the topology of a directed graph, as shown in Equation 2:

$$G =< N, N_P, E, E_P > \tag{2}$$

Among, N represent the node set; $N_p=\{Node\in N,\{(UID,hits)^n\}\}$, $n\geq 1$, record the user *UID* and the number of access nodes Node for node attribute set; E represent to the directed edges set; $E_p=\{(e\in E,\{Number\ of\ path\}^p)^m\}$, p, $m\geq 1$, for the directed edge attributes set, recorded the path to the directed edge of their number.

Definition 3: *urlID-UserID* associated matrix $M_{m\times n}$: According to the definition of a directed graph G from $G's$ node set N can get all *url* of the site, from the corresponding nodes attribute set N_p can gain access to each node's *UserID* and the corresponding access value.Thus the *urlID* as rows, the *UserID* for the column to access the number of hits for the access value, can create *urlID-UserID* associated matrix $M_{m\times n}$, as shown in Equation 3:

$$M_{m\times n}=\begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,j} & \cdots & h_{1,n} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,j} & \cdots & h_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{i,1} & h_{i,2} & \cdots & h_{i,j} & \cdots & h_{i,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{m,1} & h_{m,2} & \cdots & h_{m,j} & \cdots & h_{m,n} \end{bmatrix} \tag{3}$$

Among, $h_{i.j}$ is the number of page i *is* visited by the user j within the region for some time; each column vector $M[\cdot,j]$ indicates that the user j for all pages of the site visits; each row vector $M[i,\cdot]$ means that all user access to pages i. Therefore, row vector reflects the type of user, describes the user personalized access sub-graph; the column vector represents the site structure, and also contains user access patterns in common. Then, measuring each row vector and column vector similarity can directly get similar user groups and related Web pages, but also realize the pages and user clustering.

Definition 4: Weights associated matrix $A_{m\times n}$: Evolved from the correlation matrix $M_{m\times n}$, as shown in Equation 4:

$$A_{m\times n}=\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i,1} & a_{i,2} & \cdots & a_{i,j} & \cdots & a_{i,m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,j} & \cdots & a_{m,n} \end{bmatrix} \tag{4}$$

Among, $a_{i.j}$ is weight, $a_{i.j}=\begin{cases} 0 & \mathrm{h_{i,j}}=0 \\ 1 & \mathrm{h_{i,j}}\leq \Lambda_\mathrm{i} \\ 2 & \mathrm{h_{i,j}}\geq \Lambda_\mathrm{i} \end{cases}$, $a_{i.j}$=0 means that the user does not access the page; $a_{i.j}$=1, means that the user j interested in the contents of the page i; a_{ij}=2, means that the user j is very interested in the content of the page i; Λ_i is threshold value (threshold value according to the clustering situation may be different).

Definition 5: Similarity $sim(p_i,p_j)$:Suppose p_i and p_j are two n-dimensional space vectors, $p_i=(a_{i1},\cdots,a_{ik},\cdots,a_{im})$, $p_j=(a_{j1},\cdots,a_{jk},\cdots,a_{jm})$, $1\le k\le n$, then the similarity $sim(p_i,p_j)$ between vectors p_i and p_j shown in Equation 5:

$$Sim(p_i,p_j)=\cos(p_i,p_j)=\frac{p_i * p_j}{|p_i|\times|p_j|}=\frac{\sum_{k=1}^{n}(a_{ik}\cdot a_{jk})}{\sqrt{\sum_{k=1}^{n}(a_{ik})^2\sum_{k=1}^{n}(a_{jk})^2}} \tag{5}$$

Among, p_i*p_j means scalar product of two vectors, $|p_i|$ and $|p_j|$ are the vector norm of C and D.

3.2 Steps of Algorithm

Row vectors of *urlID-UserID* associated matrix $M_{m\times n}$ $M[i,\cdot]$ is the description of all the user on each page visit, for all users that access the Web page the same or similar circumstances must have some similarities, so they can be clustered together.

Clustering process is as follows:

Input: preprocessed log database Webdatabase;

Output: clustering results *Clu.*

Specific implementation steps are as follows:

1)*urlID-UserID* associated matrix $M_{m\times n}$ calculated threshold value for each line (the integral function), then D pretreatment. By definition 4, for any $h_{ij}>\Lambda_i$, can make $a_{ij}=2$; for any $\Lambda_i > h_{ij}>0$, can make $a_{ij}=1$; for any $h_{ij}= 0$, can make $a_{ij}=0$, thus generating the weight matrix $A_{m\times n}$.

2) By definition 5, calculate the weight matrix $A_{m\times n}$ the similarity $S_{i.j}$ between row vectors, similarity matrix $A_{m\times m}^{Sim}$ generated. In the symmetric matrix $A_{m\times m}^{Sim}$, Random $S_{i.j}\in A_{m\times m}^{Sim}$ $(1<i\le m,\ i<j\le m)$ means the similarity between i-row vector and j-row vector , the diagonal element value is 1.Because similarity matrix symmetry, just consider it part of the inverted triangle (not including diagonal). Similarity matrix as follows:

$$A_{m\times m}^{Sim}=\begin{bmatrix} 1 & \cdots & s_{1,j} & \cdots & s_{1,m} \\ & \vdots & \vdots & \vdots & \vdots \\ & & 1 & & s_{i,m} \\ & & & \vdots & \vdots \\ & & & & 1 \end{bmatrix} \tag{6}$$

3) Calculation of similarity matrix $A_{m\times m}^{Sim}$ similarity threshold value $\Lambda_.$.

$$\Lambda=\sum_{i=1}^{m}\sum_{j=1}^{m}s_{i,j} \Big/ \ A_{m\times m}^{Sim} \ \text{(Matrix } A_{m\times m}^{Sim} \text{ the number of non-zero).}$$

4) Complete the clustering operation. Output clustering results *Clu.*

3.3 Experimental Results

Selecting the logs of Liaoning Technical University of Web server (www.lntu.edu.cn) as subjects, experimental data is the data for February 11, 2010 08:22:00-15:43:21 user access the site, the entire site including the 9879 html, user access to a total of 816M, identified a total of 8217 users access services, the average transaction length is 3.7 visits. Clustering results show that the calculation of similarity Sim to take into account the sequence of the path the user to access the page, draw a high similarity to the access path. Experiment results show that the interest people in the extent of the page more reflect in the retention time of the page. Therefore, Sim calculation method proposed in this paper fully take into account the time factor on the degree of user interest, this calculation method of similarity is more practical applications, user access and description of the level of interest of the more reasonable levels. By setting the threshold value Λ_i, the number of border objects can be controlled, which effectively solve the uncertainty that Web log user access targets, and the same user may belong to several different kinds of classes, Figure 1 shows the relationship between the threshold Λ_i and the number of boundary objects.

Analysis shows that, with the threshold value Λ_i increases, the number of boundary objects in the continuous reduces, which shows that the original boundary objects with not clear classification is progressive realization classification. So that each boundary object has a defined indication, it is a effective solution to the user uncertain purpose of the browsing belong to multiple clusters of cases, and achieve stratified into categories the practical application of results.

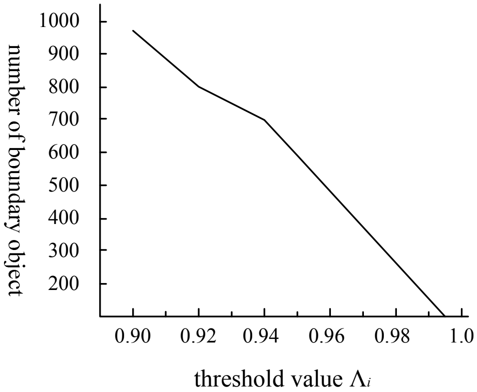


Fig. 1. Relation between number of boundary object and Λ_i

4 Conclusion

Web user access pattern mining take the user browser preference as the time reference measurement, after a choice to the next page, the longer access, the user has greater interest in, the more preference in access to, and the shorter the browsing time the opposite.

Clustering algorithm can accurately reflect the user browsing interests, without additional burden, with the advantages of simple, efficient and better system scalability. Also consider Web log preprocessing with user identification to find a class or a user browser preference path to design personalized Websites.

References

1. Barford, P., Bestavros, A., Bradley, A., et al.: Changes in Web client access patterns characteristics and caching implications. World Wide Web 2(1-2), 15–28 (1999)
2. Zhang, H., Liu, X.: A New Way to Discover User Browsing Model. Computer Applications and Software 24(2), 143–150 (2007)
3. Pa, S.K.: Web mining in soft computing framework: Relevance, state of the art and future directions. IEEE Transactions on Neural Networks (13), 223–229 (2002)
4. Fei, A.-g., Wang, X.-h.: An Information Mining Method on Web Log. Computer Application 24(6), 57–59 (2004)
5. Xu, H., Hang, J.-x., Zhu, X.-g., et al.: An Algorithm of Mining Large Reference Sequences from Web Logs. Information Technology & Informatization (1), 30–32 (2009)
6. Chen, J., Wu, J.-h.: New measuring method to predict users' browsing patterns. Computer Engineering and Applications 46(10), 209–212 (2010)

Combining Link-Based and Content-Based Classification Method

Kelun Tian

Department of Mechanical and Electrical Engineering,
Hebei Vocational & Technical College of Building Materials, Qinhuangdao, 066004, China
tklun@163.com

Abstract. Link mining is also called social network analysis. It is a new study of data mining. It is different from the traditional data mining methods. Link information is used in link mining. Link information provides richer and more accurate information about the social network. In this paper, a representation is chosen by Graph, Dyad and Subgraph for the statistical inference of mining. And then based on the defining of the graph structure and link type, the model of getting the link features is built. Last a combining link-based and content-based classification method is proposed, and this method is proved to improve the result of classification.

Keywords: link mining, link-based classification, content-based classification.

1 Introduction

The traditional data mining tasks (such as association rules mining, shopping cart analysis) view data from data set as related and independent events. Mining abundant data set, which has the structure features of heterogeneous, is an important challenge in data mining.

Link mining is also called social network analysis [1]. It is a new study of research area, which is located in multidisciplinary intersection: link analysis, Web mining [2], graph mining and social network analysis. Link mining belongs to the category of multi-relational data mining. With the deepening research of link mining, the new study task also appears, such as the prediction of the strength of link, prediction of the exist of connection, the discovery of joint quote and the mining of graph pattern, and the study of classification and clustering methods based on link of mining. All of those are hotspots of current researches.

The content of this paper is node classification method based on link. Compared with the traditional classification methods, the classification method based on link joins link information, and proposes an effective combining link-based and content-based classification method.

The rest of this paper is organized as follows: Section 1 introduces the data representation of link mining; Section 2 introduces the obtaining of link information model; Section 3 introduces combining link-based and content-based classification method; Section 4 introduces the use of data; Section 5 introduces experiments and results; and section 6 is the conclusion.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 160–168, 2011.

2 Data Representation of Link Mining

In order to accomplish the link mining task efficiently, the social network should be formalized at first. Graph provides a visual representation for formalized social network. The social network analysis proposed by Freeman has 4 basic characteristics [3]:

① The social network analysis takes more attention to the connection between the object, rather than the properties of itself.

② The information tied between the objects must be collected by systematic method.

③ Established above the graph model;

④ Meaningful information is got from these relations by using mathematics method.

Graph as a basic representation of social network is appropriate, social network is composed by the object (namely nodes of graph) and the relation tie between the object (namely the edges). Among them:

Object: Social entity, which can represent specific individuals, clubs or other social units.

Relation tie: Social entity contacts together through the relation tie. Different social networks have the different significance: the blood relationship between family members; the cooperation relationship between colleagues; Hyperlink between Web pages; Quote relationship between works, etc.

In addition to the two basic elements above, there are more complex patterns:

Dyad: Composed by two objects and the connections between them.

Subgraph: Composed by part of the social network and the connections among them. The subgraph used for analyzing characteristics of groupuscule, which belong to social networks.

Graph [4]: All objects and the relation ties among them, which used to analyze the characteristics of the whole.

Although the choice of representation is not part of the link mining, it has very important influence on statistical inference of mining. Therefore, before solving a specific social problem, the first thing is to choose a good representation. It has the more important significance for a good link mining system.

Through the introductions above, we have a basic understanding of link mining. The next is the core content of this paper: a combining link-based and content-based classification method.

3 Obtaining of Link Information Model

3.1 Definition Graph

Usually, we think the object set based on link is a directed graph essentially, node represents object, edge represents link between objects.

Definition 1
Graph: g(O, L) is the directed graph definited between O and L:

O: Represent object set. O={X_1,…, X_n}, Xi is an object or a node of graph, O is the set of nodes.

L: Represent the set of the links between objects, $L_{i\to j}$ represents a link between the object X_i and X_j. L is the set of edges.

3.2 Definition Link Types

Experience tells us the fact: users of network will always click on websites related with their own content; Thesis authors are always citing papers associated with their own theme; People always enjoy getting together with friends which have common hobbies... These give us a prompt that the content of two papers is likely to be similar if they quoted from a same paper; People who scan the same website will have the similar knowledge background extremely. Therefore, we define four basic link types as follows:

Definition 2. Link Type: The class notation of the object is a finite set {c_1,…,c_k}, c(X) denote an object X of the class c. Link Type can be divided into the following four category:

In(X_i): The in-links collection of object X_i , means the links point to X_i , {X_j | $L_{j\to i}$∈L}.

Out(X_i): The out-links collection of object X_i , means the links point to other objects, {X_j | $L_{i\to j}$∈L}.

Co-In(X_i): The joint reference collection of object X_i, {X_j | $X_j \neq X_i$,the third object X_k possess the links point to X_i and X_j }.We can consider Co-In links lke this: possess the in-links point to X_k.

Co-Out(X_i): The joint reference collection of object X_i , {X_j | $X_j \neq X_i$, X_i and X_j possess the links point to the third object X_k }. We can consider Co-Out links like this: X_i and X_j possesses the out-links point to X_k.

3.3 Feature of Links

Getting the information of links is obtained through the implementation of getting features of the links. Based on the defining of the graph structure and link type, next we need to find the pattern or model of getting the features of the links. So we tried a variety of simple mechanisms, which are based on the statistical calculations. When describing the finite data collection, statistical calculations to be more compact than storing incidence matrix .In addition, when a new object entered, these models can be adaptive , so the applied range can be more widely.

Notice that the following definition of the three models, we can get the link feature of one link type (In links, Out links, Co-In Links, Co-Out Links) at one time.

Definition 3. Model-link: {c_1,…,c_k} is the finite set of class notation , (n_1,…,n_k) is a set of vectors corresponding the class notation {c_1,…,c_k}. n_i ($1 \leq i \leq k$) denote the number of the a certain link type which between class c_i and the target object. If $n_{j,}$ $n_{i,}$ j≠I and $1 \leq j \leq k$, then the target object belongs to c_j, the result of Model-link is class c_j.

Definition 4. Count-link: {c_1,…,c_k} is the finite set of class notation , (n_1,…,n_k) is a set of vectors corresponding the class notation {c_1,…,c_k}. n_i ($1 \leq i \leq k$) denote the number of the a certain link type which between class c_i and the target object .The result of the Count-link is a set of vectors:$(n_1,\ldots,n_k)$.

Definition 5. Binary-link: $\{c_1,\ldots,c_k\}$ is the finite set of class notation , $(n_1,\ldots,n_k)$ is a set of vectors corresponding the class notation $\{c_1,\ldots,c_k\}$.If existing this kind of link type more than once between class c_i and the target object ,then n_i =1,on the contrary $n_i = 0$. The result of the Binary –link is a set of binary vectors: $(n_1,\ldots,n_k)$.

The following are examples, Fig.1:A, B, C are three classes . X is the target object.

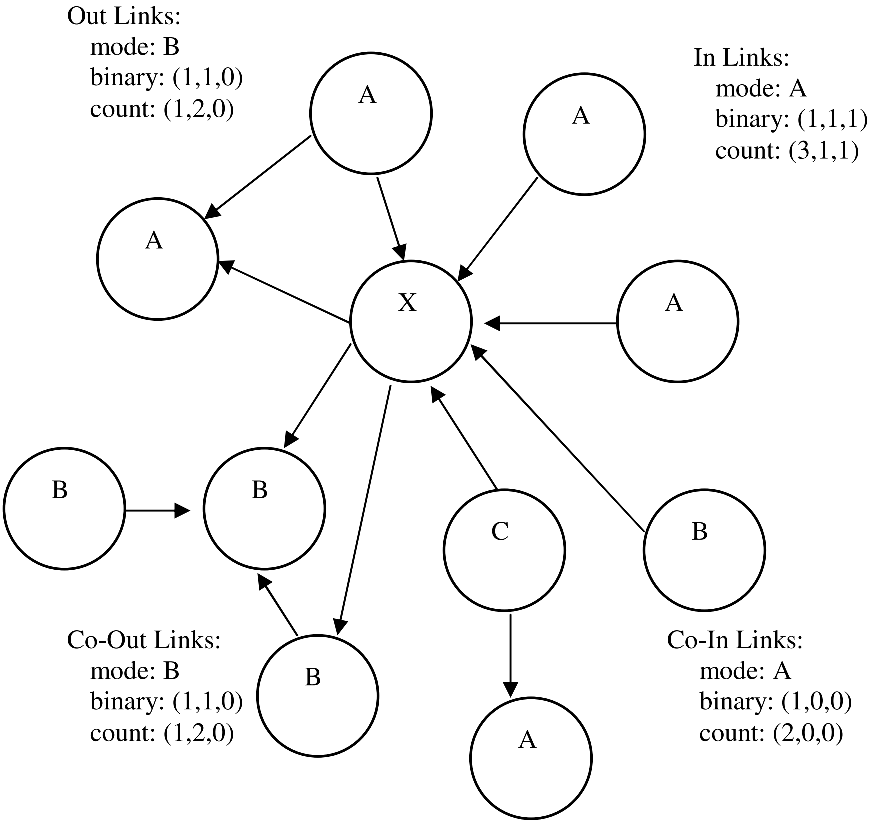


Fig. 1. Link model of object X

Fig.1 shows the link structure of target object X, and outputs the extraction results of the link feature about the three models.

It is easy to find that the Count-link provides more accurate details about the link information than Model-link and Binary-link. Count-link not only finds the classes that are associated with the target object, but also announces the proportional relationship between these classes.

4 Combining Link-Based and Content-Based Classification Method

Link mining means mining the information of the links. Link-based classification, the task is not just tap the link information, what is more important is how to improve classification results using the link information. Among the link-based classification task, data sample to be classified belongs to those who always have some contact with their class, and this contact is the hidden information of the link.

Among the link-based classification task, each data sample consists two pieces of information: link information and content information. The content information describes the properties of the samples; the link information describes the relationships between the samples. Among the tasks of link mining, the link information is more important than the content information.

Classification method of merging link information and content is divided into the following two steps:

① Dealing with the link information and content information separately (OA(X) denote the content information, LD(X) denote the link information).

② Merging the two parts by a classification model.

Traditional classification methods are based the content information, and the method has been very rich, among them the (Naïve Bayes) [5], kNN [6] and Support Vector Machine(SVM) [7] are the three most well-known methods. The classification method based on the link information is the new research direction, through the introduction of Part 2, knowing that the link information can be obtained from the implement of getting the feature of the link, in which the Count-link is considered to be the most appropriate extraction model.

In order to merge the link information and content information, we choose the Bayesian network model [8]. Bayesian network model is widely used in the information retrieval [9,10], especially for merging heterogeneous information [11].

Definition: $\{c_1,\ldots,c_k\}$ is the finite set of class notation $c_i \in \{-1, +1\}$ $(1 \le i \le k)$, given an unknown data sample X(that is ,no class notation),select Count-model to get the link feature of the sample X, then the probability of the sample X belongs to a certain class c can be expressed as:

$$P(X \mid c) = \eta(P(OA(X) \mid c)\Sigma_{t \in \{In, Out, Co\text{-}In, Co\text{-}Out\}} P(LD_t(X) \mid c)) \tag{1}$$

Where this definition, t refers to different link type, η is the normalized constant.

4.1 Content Information

$P(OA(X) \mid c)$: The probability of unknown sample X belongs to class c only considering the content information.

Expressed by function:

$$P(OA(X) \mid c_i) = class(X, c_i) \qquad (1 \le i \le k) \tag{2}$$

Where X is the unknown data sample, c_i is a certain class, class is a classification based on the content information (Native Bayes, kNN or SVM), ensuring $0 \le class(X, c_i) \le 1$.

4.2 Link Information

$$P(LD(X) \mid c) = \eta \Sigma_{t \in \{In, Out, Co\text{ - }In, Co\text{ - }Out\}} P(LD_t(X) \mid c)$$

Denote the probability of unknown sample X belongs to class c only considering the link information.

Where η is the normalized constant, ensuring:

$$0 \leq \Sigma_{t \in \{In, Out, Co\text{ - }In, Co\text{ - }Out\}} P(LD_t(X) \mid c) \leq 1$$

① Use Count-link to obtain the link feature of the unknown sample X, getting the following four vectors:

$(n_1,\ldots n_k)_{In,}$ $(n_1,\ldots n_k)_{Out}$, $(n_1,\ldots n_k)_{Co\text{-}In,}$ $(n_1,\ldots n_k)_{Co\text{-}Out}$

② Based on the link information, calculate the probability of the unknown sample X belongs to each class:

$$P(LD_t(X) \mid c_i) = \frac{n_i}{n_1 + n_2 + \ldots + n_k} \tag{3}$$

Combined (1), (2), (3), getting the probability of sample X belongs to class c_i after combining the link information and content information.

$$P(X \mid c_i) = \eta(class(X, c_i) \Sigma_{t \in \{In, Out, Co\text{ - }In, Co\text{ - }Out\}} P(LD_t(X) \mid c_i)) \ (1 \leq k \leq i) \tag{4}$$

Classification allocates unknown data sample to the biggest probability class.

Compare with structured Logistic Regression modem, the complexity of Bayesian network model reduced, when training parameters. It is important to emphasize that the belief network is a model frame, not a kind of reasoning mechanism.

5 The Use of the Data

The labeled data D^l (training set) and the unlabeled data D^u (test set) compose the data set D. In the traditional classification method, first training classifier with the labeled data, then use the classifier to distinguish the unlabeled data. But the classification method based on link is different. As we know, the relationship in link is dynamic. When a classified numbered sample was joined in the labeled data successfully, it will bring new changes to the link information of the labeled set.

First, create Bayesian network model for the labeled data set D^l (training set). Then classify the unlabeled data set D^u (test set). Finally train the new model with the classified unlabeled data and the original labeled data set.

It can be classified into the following two steps:

Step1: initialization. Creating Bayesian network model and training classification according to the information of content and link only use the labeled data set.

Step2: iterative (usually use EM-like iterative algorithm)

1. Classify the unlabeled data use the classifier that is trained in step1.

2. Recalculating the link features of the labeled data set sample. Reevaluate the Bayesian network model parameters.

By iteration and training all the unlabeled data classified successfully, unlabeled data and labeled data are used in the process of training classifier and parameter.

6 Experiment and Result

Experimental data from Cora [12] data set(a data set build by a CiteSeer); WebKB [13] data set(World Wide Web).

Establish the experiment data set: CoraⅠ, CoraⅡ, SubWebKB (part data from Cora and WebKB).

6.1 Define Experimental Metrics

TP (true positive): sample x belonging to class c_i, through the classifier identify, it assigned to class c_i indeed.

FP (false positive): sample x not belonging to class c_i, but after identify by classifier, it assigned to the class c_i.

TN (true negative): sample x not belonging to c_i, through the classifier identify, it assuredly did not belonging to c_i.

FN (false negative): sample x belonging to c_i, but after identify by classifier, it did not belonging to c_i.

Accuracy is the percentage of the properly labeled sample, (TP+TN) / (TP+TN+ FP+FN).

Precision, be marked in the same kind of samples, the percentage of correct marked TP / (TP+FP).

Recall, belongs to the same kind of samples and the percentage of be properly labeled TP/ (TP+FN).

6.2 Experimental Results

The experiments show that combined link information and content information is superior than based on content only. The accuracy of Count model is better than Mode model, and they all better than Binary mode. The reason is the mode which is based on linking use a lot of adjacency sample attribute, Mode and Count model make full use of these information. In a sense, Binary model only describe the kind of existence, but lost the information about frequency. Especially in the reference field, the advantage of Count model is more apparent.

Such as table 1, table 2, table3, three conditions are compared in Cora I, Cora II and SubWebKB: the result based on only consider the content; the result based on only consider Link information; the result after combine the content and the link information.

Table 1. Classified re report table on Cora I data set

	CoraⅠ		
	Content	Links	Combine
Avg accuracy	66.39	78.16	82.27
Avg precision	68.07	79.11	85.26
Avg recall	63.25	75.76	81.41

Table 2. Classified report table on Cora Ⅱ data set

	Cora Ⅱ		
	Content	Links	Combine
Avg accuracy	62.98	79.96	81.57
Avg precision	65.48	78.46	80.29
Avg recall	48.37	75.26	77.34

Table 3. Classified report table on SubWebKB data set

	SubWebKB		
	Content	Links	Combine
Avg accuracy	82.12	56.44	88.39
Avg precision	72.34	68.81	75.96
Avg recall	78.91	59.33	80.82

Experimental results show that model of combing link-based and content-based classification is better. But compare with only consider the content or link information, it show different results. When consider link information only, the results is better than consider content only based on Cora I data set and Cora II data set. But, based on SubWebKB data set, the result is opposite, and the result is not very ideal when only consider the link information. The reason is that content can supply better initialization entrance. It compensated the lack content when data focused on fully linked information. When link information rare of focused data, the performance maybe bad.

Comparing the contribution of different type of optimization classification, it shows that all link type is good. But when consider one part, Out Links and Co-In Links seems supply more information, and when based on SubWebKB is not good.

Table 4. Accuracy classified table labeled only and complete

	Labeled links only	Complete Links
Cora I	68.72	82.27
Cora Ⅱ	70.36	81.57
SubWebKB	85.66	88.39

Table 4 Comparing the accuracy of classified results only use labeled data set and unlabeled data set.

The classification results improved by use the increase unlabeled data set. Link information may improve the classified results, in terms of , we can change the error results based on content only according to the link information .And how to coordinate the percentage between link information and content information in the process of classify will be a potential research direction.

7 Conclusions

Many data sets have complex structure, and something is always interacting in reality. Link mining is the best method in digging mass structured data. Mining link information and using model obtain information is the core task of link mining .In this paper, we focused on using the link information improve the classification results.

This article put forward a kind of simple framework based of statistical method in order to mode the future of link. Bayesian network model played a very good effect role when merge the link information and the content information. It found four link types, analyzed their importance, and added the unlabeled data to the process of training classifier, and discussed the relationship between link information and content information. This content will provide important base for future research.

References

1. Han, J.W., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2000)
2. Chakrabarti, S.: Mining the Web. Morgan Kaufman, San Francisco (2000)
3. Freeman, L.: The Development of Social Network Analysis. In: A Study in the Sociology of Science 2004. Empirical Press, Vancouver (2004)
4. Cook, D., Holder, L.: Graph-based data mining. IEEE Intelligent Systems and Their Applications 15(2), 32–41 (2000)
5. Mccallum, A., Nigam, K.: A comparison of event models for Naïve Bayes text classification. In: Proceeding of AAAI/ICML-1998 Workshop on Principles of Database Systems, pp. 41–48 (1998)
6. Kantarcioglu, M., Clifton, C.: Privacy preserving k-nn classifier. In: ICDE (2005)
7. Osuna, E., Freund, R., Girosi, F.: An improved training algorithm for support vector machines. In: Neural Networks for Signal Processing VII–Proceedings of the 1997 IEEE Workshop, pp. 276–285 (1997)
8. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible inference. Morgan Kaufmanns, San Francisco (1988)
9. Ribeiro, N.B., Muntz, R.: A belief network model for IR. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, pp. 253–260 (1996)
10. Turtle, H., Croft, W.B.: Evaluation of inference network-based retrieval model. ACM Transactions on Information Systems 9(3), 187–222 (1991)
11. Ribeiro, N.B., Silva, I., Muntz, R.: Soft Computing in Information Retrieval: Techniques and Applications. In: Chapter 11-Bayesian Network Models for IR, ch. 11, pp. 259–291. Springer, Heidelberg (2000)
12. Mccallum, A., Nigam, K., Rennie, J., et al.: Automating the Construction of Internet Portals with Machine Learning. Information Retrieval 3(2), 127–163 (2000)
13. Craven, M., Dipasquo, D., Freitag, D., et al.: Learning to extract symbolic knowledge from the world wide web. In: Madison (ed.) 15th Conference of the American Association for Artificial Intelligence 1998, pp. 509–516. AAAI Press, Menlo Park (1998)

Chinese Expert Entity Homepage Recognition Based on Co-EM*

Li Liu[1], Zhengtao Yu[1,2], and Lina Li[1]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650051, China
[2] Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming 650051, China
{liuli0407,ztyu}@hotmail.com

Abstract. Focused on the problem of numerous labeling works on the expert homepage in the procedure of Chinese expert entity homepage recognition, in this paper, a method of Chinese expert entity homepage recognition based on the Co-EM proposed. In detail, firstly, collect the names of Chinese expert entity and the corresponding web pages, and then label a small quantity of web pages. Secondly for Chinese entity characteristics, extract the hyperlink features and the web page content features as two independent feature sets. Thirdly train the hyperlink classifier using the hyperlink feature set and label the all the expert entity homepages, and then train the content classifier using the web page content feature set and the labels which were labeled by the hyperlink classifier. Use the labels which were labeled by the content classifier to update the hyperlink classifier. Repeat the procedure until the two classifiers converge. Finally, experiments were done by employing the method of 10-fold cross validation. The results show that the method based on the Co-EM semi-supervised algorithm can uses the unlabeled web pages effectively and there is an increase of accuracy of recognition compared with using the labeled web pages only.

Keywords: Chinese Expert Entity Homepage Recognition, Semi-supervised learning, Co-EM.

1 Introduction

With the increase demand of expert's information, how to find out the expert entity homepages from a mass of Internet web pages quickly and accurately has gradually become people's strong demand. Given an expert entity name, how to find the expert's homepage and detail information directly has an important prospect.

In the English expert entity homepage recognition, the famous international conference TREC [1] (Text Retrieval Conference) set the Enterprise Search task from

* This paper is supported by National Nature Science Foundation (60863011), Yunnan Nature Science Foundation (2008CC023), Yunnan Young and Middle-Aged Science and Technology Leaders Foundation (2007PY01-11).

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 169–177, 2011.

2005, besides, it set the English Expert Search as its sub-task. In the early 90s of 20 century, Hewlett-Packard developed the CONNEX system which is used to find homepages of the employees [2]. Since then, Dawit Yimam [3] proposed a DEMOIR (Dynamic Expertise Modeling from Organizational Information Resources) approach to carry out the research on entity homepage recognition. Campbell et al. [4] also analyzed the link structure defined by authors and receivers of emails using a modified version of the Hyperlink-Induced Topic Search (HITS) algorithm to identify authorities and return information and homepages of experts. In addition, probabilistic approaches presented by Macdonald and Ounis [5] and formal methods proposed by Balog et al. [6] have been developed based on the assumption to search experts and the corresponding homepages. Lina Li et al. [7] defined the entity features related to the features of link and web page content, and then used the AdaBoost algorithm to train the classifier. Yi Fang, Luo Si et al. [8, 9, 10] also developed a search engine used for the Indiana Database of University Research Expertise (INDURE) to obtain the specific information and homepages of experts. Meanwhile, the SPUD (Skills Planning and Development) project [11] of Microsoft and the SmallBlue [12] System of IBM both are the typical applications in this area. In the Chinese expert entity recognition, Lina Li et al. [13] adopted the method which combines the feature of link and web page content and the J48 algorithm and the accuracy rate reaches 81.05%. The Chinese expert entity recognition can be transformed into the classification task. But collecting and labeling a mass of training data is necessary in the procedure of machine learning, in the real world application, labeling a mass of web pages manually is costly, but we can get a mass of unlabeled web pages through the search engine is easy. Therefore, the semi-supervised learning of how to combine labeled data with unlabeled data becomes the major concern of expert entity homepage recognition researches.

Various semi-supervise learning methods have been proposed in the machine learning domain. Co-training introduced by Blum and Mitchell [14] is a semi-supervised learning paradigm, which trains two separate classifiers with the labeled data and let them label the unlabeled examples with high confidence for each other. Each classifier is retrained with the additional training examples given by the other classifier. Kamal Nigam [15] proposed the Co-EM algorithm based on the Co-training and EM theory which solves the problem of “noisy data” when we used the Co-training algorithm to train two classifiers and used the two classifier to label data in the same time.

In this paper, with the help of search engine, we collected a mass expert homepage data and labeled a little data. And then for Chinese entity characteristics, defined the expert entity features related to the features of link and web page content as two independent feature sets. Finally we trained the expert homepage classifier by the Co-EM algorithm. The result of experiment shows that the classifier which is trained by the algorithm we proposed can use the unlabeled data effectively and improve the accuracy rate.

The organization of this paper is as follow: In section 2, we describe the characteristics of Chinese expert homepage and the method of extracting the features; in section 3, the method of Chinese expert entity recognition combines Co-EM algorithm is introduced; in section 4, we describe the result of Chinese expert entity recognition and analyze the result. Finally, some concluding remarks are given in section 5.

2 Feature Extraction in Expert Homepage

2.1 Collection of Chinese Expert Entity Homepage

In this paper, we collected the random expert entity names from several universities' websites, and then input the names into search engine baidu (http://www.baidu.com) to get the top 5 web pages corresponding to every name. Next, we mark the order of the web pages in the search engine and labeled a little web page. In detail, 0 denotes the web page is not the expert's homepage, 1 denotes the web page is the expert's detail information but isn't the homepage, 2 denotes the web page is expert's homepage and 99 denotes the web page is unlabeled. And then, the expert entity web pages can be described as a six-tuple SEU: $SEU = (ID, EN, Content, Rank, Url, Score)$. In which, ID is the order number when we collected the web page, EN is the expert entity name, Content is the content of the web page, Url is the hyperlink corresponding to the web page, Rank is the web page's rank number in the search engine's list and the score denotes the type of the web page.

2.2 Characteristics of Chinese Expert Homepage

In general, there are a mount of features which is relevant to the expert's homepage in the web page's hyperlink. From the analysis of the hyperlinks of the Web Pages, we can see that, there are several special characters such as "?", "/", "=" et al, besides, the web page's rank in the search engine is high. We also discovered that: the hyperlink which has the words such as "news","baike" is not the expert's homepage's hyperlink; the expert's homepage's hyperlink has fewer characters like "%" et al, a mount of expert's homepage's hyperlink has several numbers. So for the characters of expert's homepage's hyperlinks, we adopt 14 typical features (which are shown as feature 1-14 in table 1 in subsection 2.3).

Besides, the content of the web page also has several features can be used for the expert entity recognition. For example: there are several words such as "姓名(name)", "研究(research)", "论文(papers)", "教授(professor)", "学院(college)", "报告(report)", "大学(university)" corresponding to the expert entity name, the hyperlink and the images in the web page content both are the typical features of expert homepage. We adopted 4 typical features in the web page's content(which are shown as feature 15-18 in table 1 in subsection 2.3), besides, borrowed ideas from the text classification, we remove the html tags of the web pages , segment the words , remove the stop-words, and then calculate the information gain(IG) of every word. The equation of IG is shown as equation 1:

$$G(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t) \\ + P(\bar{t})\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t}) \quad . \tag{1}$$

In equation 1, t denotes a term in one question, $\{c_i\}_{i=1}^{m}$ denotes the classes of the target space. $P(c_i)$ is the probability of class c_i occurs in the train corpus. $P(t)$ is the probability of term t occurring in the train corpus. $P(c_i | t)$ denotes the co-occurrence probability of c_i and t in the train corpus. $P(c_i | \bar{t})$ is the probability of t not occurring in the train corpus. After the calculation, we adopt the top 10 words which have the higher IG values to extend the web page content features

2.3 Feature Exaction of Chinese Expert Homepage

After the analysis of subsection 2.2, in this paper, we adopted the hyperlink features and web content features as two independent feature sets separately. The feature list of the Chinese expert homepage is shown as table 1:

Table 1. The feature list of Chinese expert homepage

Feature id	Feature Description	type
Feature1	Whether the hyperlink include “news”	Num1 {0,1}
Feature2	The length of the hyperlink	Num2 integer
Feature 3	The count of “/” in the hyperlink	Num3 integer
Feature 4	The count of “?” in the hyperlink	Num4 integer
Feature 5	The baidu rank of the homepage	Num5 integer
Feature 6	The count of “%” in the hyperlink	Num6 integer
Feature 7	The count of “=” in the hyperlink	Num7 integer
Feature 8	The count of “_” in the hyperlink	Num8 integer
Feature 9	Whether the hyperlink include “baike”	Num9 {0,1}
Feature 10	Whether existing whole numeric information between two “/ “ in HYPERLINK, such as “/888/”	Num10 {0,1}
Feature 11	Whether the hyperlink include “~”	Num11 {0,1}
Feature 12	The count of numbers in the hyperlink	Num12 integer
Feature 13	The count of numbers after the last second “/” of the hyperlink	Num13 integer
Feature 14	The count of “&” in the hyperlink	Num14 integer
Feature 15	The count of entity name in the web page content	Num15 integer
Feature 16	The count of images in the web page content	Num16 integer
Feature 17	The count of hyperlinks in the web page content	Num17 integer
Feature 18	Whether the web page content include the words “姓名(name)、研究(research)、学会(academy)、论文(papers)、教授(professor)、学院(college)、报告(report)、大学(university)、协会(society)、学部(faculty)、研究所(institute)、博士(PhD)、专门(specialist)、所属(belonged to)、讲座(lecture)、课题(topic)、发表(publish)、出版(publishing)、领域(area)、项目(project)”	Num18 integer
Feature 19-28	Term1-Term10 which were adopted by calculating the Terms’ IG values	Num19-28 integer

In the table 1, the feature 1-14 are the features based on the homepage’s hyperlinks, the feature 15-28 are the features based on the homepage’s content. Focused on the six-tuple SEU we get from the search engine in subsection 2.1, from the definitions in the table 1, we adopted the feature 1-14 as hyperlink feature set and adopted the feature 15-28 as the content feature set.

3 Chinese Expert Homepage Recognition Based on Co-EM

3.1 Modified Bayesian Model

Bayesian Model is a simple but effective text classification algorithm for learning from labeled data. The model c calculates the probability of a document belongs to a class. We choose the Modified Bayesian model as the fundamental classifier. In details, the algorithm describe as follows:

Suppose that the training data is consist of features, $c_j \in C = \{c_1, \cdots, c_{|C|}\}$ denotes the jth class, and then we can get the probability of document D_i belongs to c_j which is shown as Equation 2

$$P(c_j \mid D_i) = P(c_j \mid w_1, \cdots, w_n) = \frac{P(c_j, w_1, \cdots, w_n)}{P(w_1, \cdots, w_n)} \tag{2}$$

In the equation 2, w_i is the feature of the document D_i and the denominator is a constant when the training data is determined. The Modified Bayesian model supposes the features of the document are independent, so equation 3 is gained because of the independent probability.

$$P(c_j, w_1, \cdots, w_n) = P(c_j, w_1) \times P(c_j, w_2) \times \cdots \times P(c_j, w_n) \tag{3}$$

After deal with the weight of TF-IDF, the equation (4) is obtained:

$$P(c_j, w_t) = \frac{0.5 + n(c_j, w_t)}{V + \sum_{i=1}^{V} n(c_j, w_t)} \times \log \frac{V + 0.1}{M + 0.1} \tag{4}$$

In which $n(c_j, w_t)$ denotes the number of times feature w_t appears in the class c_j . $\sum_{i=1}^{V} n(c_j, w_t)$ denotes the total number of the times feature w_t appears in all the classes. V is the count of classes, M denotes the feature w_t appears in M classes, the 0.5 is smooth factor.

At last, we can predict a document's label by maximizing the equation 5

$$\arg\max P(c_j \mid q_i) \propto \arg\max P(c_j, w_1, \cdots, w_n) \propto \arg\max(P(c_j, w_1) \times \cdots \times P(c_j, w_n)) \tag{5}$$

3.2 Co-EM Algorithm

The Co-EM algorithm is a semi-supervised algorithm which combines EM theory and the Co-training theory. The algorithm's procedure is similar with the Co-training, first train a classifier C_1 using a independent feature set, then use C_1 labels all the unlabeled data and use the other independent feature set and the labels which are labeled by C_1 to train the classifier C_2, finally use C_2 labels all the data and use the labels which are labeled by C_2 to update the parameters of C_1, repeat the procedure until the two classifiers converge. The algorithm is described as the following:

Algorithm 1. The Co-EM algorithm

```
Input: labeled training set $Q_l$, unlabeled training set $Q_u$, number of iterations K
Output: Two classifiers, $C_1$ and $C_2$. These predictions combined by multiplying together
Process:
For t=1 to K do
    Train the naive bayes classifier $C_1$ using the hyperlink features of $Q_l$ only.
    Predict the unlabeled set $Q_u$ using $C_1$ and train the other modified bayes classifier $C_2$ using the content features of $Q_l$ and $Q_u$
    Predict $Q_u$ using $C_2$ and use the labels which were predicted by $C_2$ to update the parameters of $C_1$.
End for
```

The major difference between the Co-training and Co-EM is that, the Co-EM does not use the first classifier to classify the unlabeled data directly but use the posterior to classify the unlabeled data; the labels of the data can be changed in the following training procedure. In the Co-training algorithm, the original classifier's accuracy rate has a big influence on the final accuracy rate. And the Co-EM algorithm solves the problem.

3.3 Chinese Expert Homepage Recognition based on the Co-EM

We get the Chinese expert entity name from the universities' website at random, retrieve a mass of web pages from the search engine, and then label a little manually and build the six-tuple SEU. the data set marked as $Q = Q_u \cup Q_l$, in which Q_u is the unlabeled web page, Q_l is the labeled web page. Next, we train the Modified Bayesian model C_1 using the hyperlink feature set, then label the Q_u by C_1,train the other modified Bayesian model C_2 using the content feature and labels which are labeled by C_1, update the classifier C_1 using the labels which are labeled by C_2, repeat the procedure and the number of times of repeating is 10. Finally we get two classifiers C_1 and C_2, and for the prediction of a new data, we maximum the $p(c_1) \times p(c_2)$, in which p (C_1) is the probability of classifier C_1 and the probability of classifier C_2.

4 Experiment and Analysis

The experimental data is the expert entity names which are obtained from the universities' website at random, and then with the help of search engine, 2113 Chinese expert web pages are obtained. Next label 300 web pages manually, that is to assign the value of Score of the six-tuple SEU=(ID, EN, Content, Rank, Url, Score). In order to evaluate the effectiveness of Co-EM, In this paper, the methods of feature extraction are choose as hyperlink feature, content feature, hyperlink feature and the content feature; the methods of classification are choose as SVM, Modified Bayes, J48, Co-EM.

Experiment uses the method of 10-fold cross validation to verify the classification results with different classifiers. The specific classification results are shown in Table 2.

Table 2. The accuracy rate of different methods of feature extraction and classification

Feature	Accuracy			
	SVM	BAYES	J48	Co-EM
HYPERLINK	65.54%	75.49%	80.17%	_
CONTENT	63.12%	73.83%	77.84%	_
HYPERLINK+CONTENT	66.21%	77.61%	82.49%	83.48%

Note: The Co-EM algorithm can not be used only with the hyperlink feature or the content feature, so there is no accuracy rate in the table when used the Co-EM

The result shows that: In the supervised learning, the accuracy rate of the feature extraction method which combines the hyperlink feature and the content feature is higher than the method which only uses the hyperlink feature or content feature. In which J48 method, combined with the feature of hyperlink and the content obtain a better result, the accuracy rate reaches 82.49%. In the semi-supervised learning, using the Co-EM algorithm, the accuracy rate reaches 83.48%, it proves that, Co-EM algorithm can use the unlabeled data effectively and reduce the labeling work; meanwhile, an increase of classification accuracy rate is obtained.

In order to evaluate the convergent speed of Co-EM algorithm in the iterative procedure, the classification accuracy of 10 iterative procedures are shown as Figure 1.

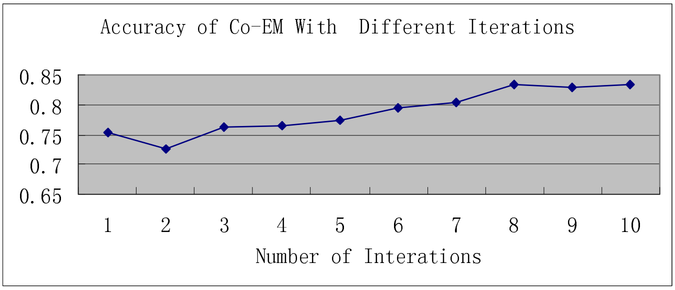


Fig. 1. The classification accuracy of Co-EM algorithm in 10 iterative procedures

The results show that: in the previous iteration, the accuracy of classification is low. After several iterations, the accuracy of classification increases and changes to steady, in the figure 1, after 8 iterations, there is a little change in accuracy of classification. Therefore, Co-EM algorithm has a quick speed convergent speed; that is this model has a quick speed of training.

In order to evaluate the influence of labeled data in the Co-EM algorithm, in this paper, we add some labeled data into the training procedure. The labeled web pages are 350,400,450,500,550,600; the number of iterations is 10. The result is shown in Figure 2.

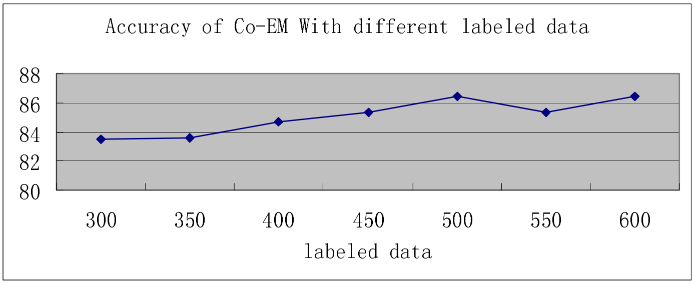


Fig. 2. The classification Accuracy of different count of labeled data

The result shows that: when the labeled data increases in the range of 350-500, the accuracy increases, and when the labeled data reaches 500, there is a little change in accuracy of classification. It shows that, in the procedure of Chinese expert homepage recognition, we can obtain a steady accuracy without labeling a mass of web pages; the Co-EM algorithm can use the unlabeled data effectively, reduce the work of labeling and improve the accuracy of classification.

5 Conclusion

Chinese Expert Entity Homepage Recognition is an important component of expert system, with the help of semi-supervised learning algorithm, there is less work on labeling data manually. The result of experiment shows that the Co-EM algorithm uses the unlabeled data effectively and improves the accuracy of classification. In future, we will focus on the feature optimization with hyperlink and web page and the optimal homepage learning algorithm improvement.

References

1. http://trec.nist.gov/
2. Davenport, T.: Knowledge management at Hewlett Packard, Center for Business Innovation (1996)
3. Yimam-Seid, A.: Expert finding systems for organizations: Problem and domain analysis and the DEMOIR approach. Journal of Organizational Computing and Electronic Commerce 13(1), 1–24 (2003)

4. Campbell, C.S., Maglio, P.P., Cozzi, A., et al.: Expertise identification using email communications. In: CIKM 2003: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 528–531. ACM Press, New York (2003)
5. Macdonald, C., Ounis, I.: Voting for candidates: adapting data fusion techniques for an expert search task. In: CIKM 2006: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 387–396 (2006)
6. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, USA, pp. 43–50 (2006)
7. Li, L., Yu, Z., Zou, J., Su, L., Xian, Y., Mao, C.: Research on the Method of Entity Homepage Recognition. Journal of Computational Information Systems (2009)
8. Fang, Y., Si, L., Mathur, A.: FacFinder: Search for Expertise in Academic Institutions, Technical Report, SERC-TR-294 and Department of Computer Science, Purdue University (2008)
9. Fang, Y., Si, L., Mathur, A.: Learning to Rank Expertise Information in Heterogeneous Information Sources. In: SIGIR 2009 Workshop on Learning to Rank for Information Retrieval (SIGIR Workshop), Boston, USA (July 2009)
10. `http://www2.itap.purdue.edu/indure/`
11. Davenport, T., Prusak, L.: Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press, Boston (1998)
12. Lin, C., Griffiths-Fisher, V., Ehrlich, K., Desforges, C.: SmallBlue: People Mining for Expertise Search and Social Network Analysis. IEEE Multimedia Magazine (2008)
13. Li, L., Yu, Z., Wang, Y., Mao, C., Guo, J.: Research on the Method of Chinese Expert Entity Homepage Recognition. Journal of Guangxi Normal University(Natural Science Edition) (March 2011)
14. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th Annual Conference on Computational Learning Theory, Wisconsin, MI, pp. 92–100 (1998)
15. Nigam, K., Ghani, R.: Analyzing the effectiveness and applicability of co-training. In: Ninth International Conference on Information and Knowledge Management, pp. 86–93 (2000)

Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers

Xin Wang[1], Chaofei Wang[2], and Junyi Shen[1]

[1] Department of Electronic and Information Engineering
Xi'an Jiaotong University, Xi'an 710049, China
wangx@cetin.net.cn
[2] China Defense Science and Technology Information Center
Beijing 100142, China
wcf-119@163.com

Abstract. Semi-supervised clustering uses a small amount of labeled data to aid and bias the clustering of unlabeled data. This paper explores the usage of labeled data to generate and optimize initial cluster centers for k-means algorithm. It proposes a max-distance search approach in order to find some optimal initial cluster centers from unlabeled data, especially when labeled data can't provide enough initial cluster centers. Experimental results demonstrate the advantages of this method over standard random selection and partial random selection, in which some initial cluster centers come from labeled data while the other come from unlabeled data by random selection.

Keywords: semi-supervised clustering, k-means, initial cluster centers, max-distance search.

1 Introduction

In pattern recognition, machine learning and relative fields, supervised learning is a method based on a great amount of labeled data to supply training sets. However, there are a mass of unlabeled data compared with limited labeled ones in many practical problems, such as Web Page Classification, Gene Analysis, Voice Recognition and so on. It is a truth that limited training data can not provide sufficient distribution information of dataset, which leads to unsatisfied result in practical applications. Meanwhile, significant number of unlabeled data is useless in supervised learning. Unsupervised learning method tries to build classifier by means of detecting hidden structures in unlabeled data. But it is difficult to make sure the accuracy on dealing with mass data. Therefore, it begins to raise more concern and become a new hot research issue that semi-supervised learning method utilizes comprehensively a bit labeled data and massive unlabeled ones.

For different learning tasks, semi- supervised learning can be divided into semi-supervised classification and semi-supervised clustering [1]. Semi-supervised classification [2, 3] makes use of a large amount of unlabeled data to enlarge the training set, which can compensate the disadvantage due to the inadequate labeled

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 178–187, 2011.

data. Semi-supervised clustering [4-7] utilizes some labeled data to obtain a better clustering. This paper explores the use of labeled data to generate initial cluster centers for k-means algorithm, which biases clustering towards a good direction. Semi-supervised classification algorithms cannot instead of semi- supervised clustering algorithms to complete some learning tasks, in which a small amount of labeled data can not reflect the structure of dataset [8].

This paper introduces a semi-supervised k-means clustering algorithm, called Seeded K-means (Basu etc., 2002), which uses labeled data to generates initial cluster centers. But this algorithm is proceeded in the assumption that the labeled dataset covers all categories (in other words, each category has one labeled data at least [9]). It is obvious that this hypothesis is very limited. A more common problem is that we only have some labeled data which come from a part of categories, and we have not any one labeled data which come from the other categories. Based on this situation, this paper presents a max-distance search approach to find some optimal initial cluster centers from unlabeled data. We present results of experiments which demonstrate the advantages of our method over standard random selection and partial random selection, in which some initial cluster centers come from labeled data while the other come from unlabeled data by random selection.

2 Problem Description

In this section, we definite some concepts used in this paper, and then introduce the background knowledge, finally describe the problem we are facing.

2.1 Seed Set

Given a dataset X as previously mentioned, k-means clustering of the dataset generates a k-partitioning $\{X_l\}_{l=1}^{k}$ of X . Let $S \subseteq X$, called the Seed Set, be a subset consisting of labeled data (called the Seeds) which are provided as follows: for each $x_i \in S$, the user provides the cluster X_l of the partition to which it belongs. The Seed Set is used to generate initial cluster centers for k-means algorithm.

2.2 Complete Seed Set and Incomplete Seed Set

We assume that corresponding to each partition X_l of X , there is typically at least one seed $x_i \in S$. Note that we get a Seed Set with a k-partitioning $\{S_l\}_{l=1}^{k}$, called the Complete Seed Set.

However, in most cases we can't get the Seed Set with a k-partitioning $\{S_l\}_{l=1}^{k}$ but a j-partitioning $\{S_l\}_{l=1}^{j},(j<k)$, called the Incomplete Seed Set, which means that corresponding to some partitions of X (actually j of k), there is typically at least one seed $x_i \in S$, but corresponding to the other partitions (actually $k-j$ of k), there is not one seed $x_i \in S$.

An example of Complete Seed Set and Incomplete Seed Set is presented in Fig.1.

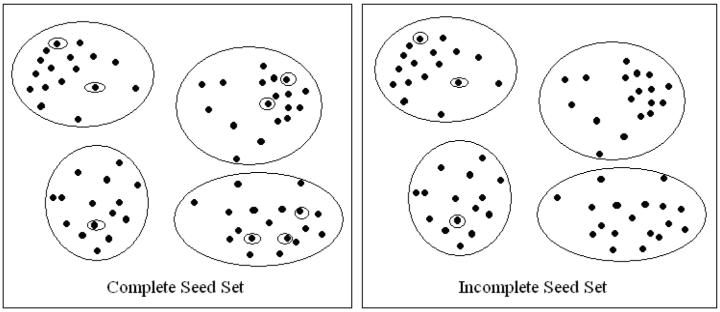
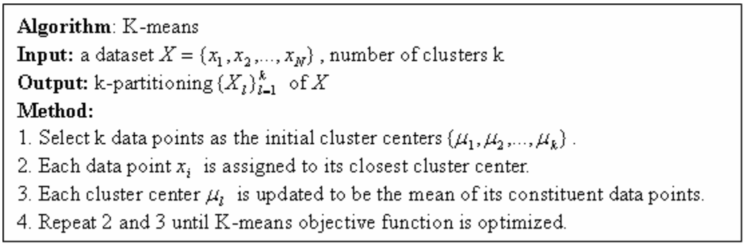


Fig. 1. Complete Seed Set and Incomplete Seed Set

2.3 Semi–supervised K-Means Clustering

K-means clustering (MacQueen, 1967) is a method commonly used to automatically partition a dataset into k clusters. The algorithm is presented in detail in Fig.2 [9, 10].



Algorithm: K-means
Input: a dataset $X = \{x_1, x_2, ..., x_N\}$, number of clusters k
Output: k-partitioning $\{X_l\}_{l=1}^{k}$ of X
Method:

1. Select k data points as the initial cluster centers $\{\mu_1, \mu_2, ..., \mu_k\}$.
2. Each data point x_i is assigned to its closest cluster center.
3. Each cluster center μ_l is updated to be the mean of its constituent data points.
4. Repeat 2 and 3 until K-means objective function is optimized.

Fig. 2. K-means Algorithm

It is well known that the most challenge of k-means algorithm is selection of the initial cluster centers. The traditional k-means algorithm randomly selects k data points as initial cluster centers from unlabeled dataset, which leads to the chances of it getting stuck in poor local optima. In the research of semi-supervised k-means algorithm, the outbreak is taking advantage of labeled data to obtain initial cluster centers.

So the problem is how can we utilize the limited labeled data to obtain good initial cluster centers?

3 Algorithms

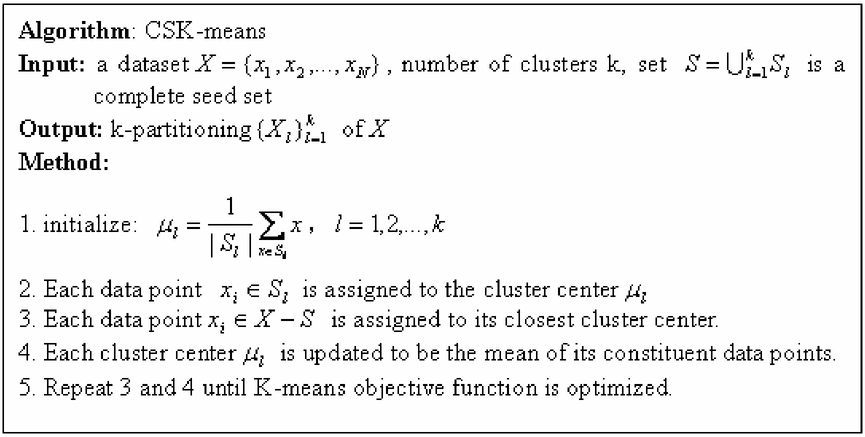
In this section, we explain how semi-supervision can be incorporated into the k-means algorithm by optimizing initial cluster centers. According to the two kinds of Seed Set mentioned in chapter 2.2, we can divide semi-supervised k-means into the semi-supervised k-means based on Complete Seed Set, called “CSK-means” for short and the semi-supervised k-means based on Incomplete Seed Set, called "ISK-means" for short, and then we give the mathematical motivation behind the two proposed algorithms.

3.2 CSK-Means

In CSK-means, a complete seed set is used to initialize the k-means algorithm. Thus, rather than initializing k-means from k random data points, the initial center of the lth cluster is initialized with the mean of the lth partition S_l of the seed set. The calculation of the initial cluster centers $\{\mu_l\}_{l=1,2,\ldots,k}$ can be written as:

$$\mu_l = \frac{1}{|S_l|}\sum_{x \in S_l} x, l = 1,2,\ldots,k \tag{1}$$

The algorithm is presented in detail in Fig. 3.



```
Algorithm: CSK-means
Input: a dataset $X = \{x_1, x_2, \ldots, x_N\}$, number of clusters k, set $S = \bigcup_{l=1}^{k} S_l$ is a
        complete seed set
Output: k-partitioning $\{X_l\}_{l=1}^{k}$ of $X$
Method:
1. initialize: $\mu_l = \frac{1}{|S_l|}\sum_{x \in S_l} x$, $l = 1,2,\ldots,k$
2. Each data point $x_i \in S_l$ is assigned to the cluster center $\mu_l$
3. Each data point $x_i \in X - S$ is assigned to its closest cluster center.
4. Each cluster center $\mu_l$ is updated to be the mean of its constituent data points.
5. Repeat 3 and 4 until K-means objective function is optimized.
```

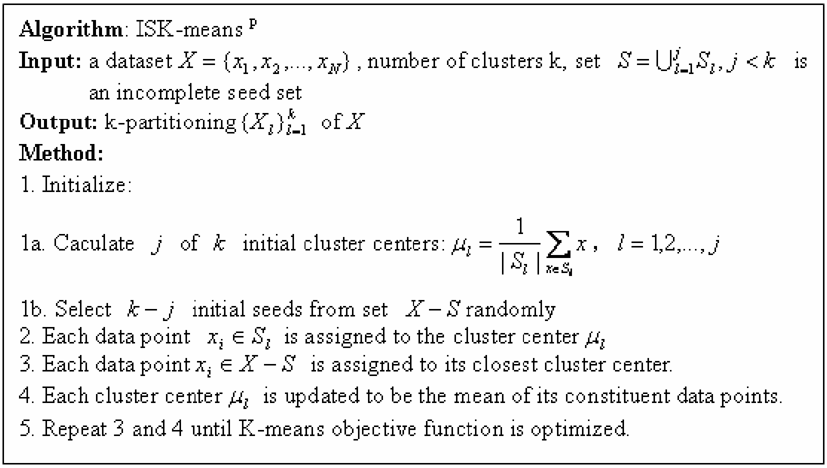
Fig. 3. CSK-means clustering algorithm

3.3 ISK-Means

In ISK-means, an incomplete seed set is used to initialize the k-means algorithm. Thus, rather than CSK-means, the key point of ISK-means is how to calculate the $k-j$ initial cluster centers when j of k initial cluster centers can be initialized with the means of the j partitions $\{S_l\}_{l=1}^{j}, (j < k)$.

Partial random selection. A simple method to initialize the $k-j$ initial cluster centers is selecting them randomly from the unlabeled dataset, which is called partial random selection. We can call it ISK-means ^{P}, short for the ISK-means algorithm with the method of partial random selection. The algorithm is presented in detail in Fig. 4.

Easy to analysis, this algorithm has a better performance than unsupervised k-means, but the method of partial random selection still can make it get stuck in poor local optima.



```
Algorithm: ISK-means P
Input: a dataset $X = \{x_1, x_2, ..., x_N\}$, number of clusters k, set $S = \bigcup_{l=1}^{j} S_l, j < k$ is
       an incomplete seed set
Output: k-partitioning $\{X_l\}_{l=1}^{k}$ of $X$
Method:
1. Initialize:
1a. Caculate $j$ of $k$ initial cluster centers: $\mu_l = \frac{1}{|S_l|}\sum_{x \in S_l} x$, $l = 1,2,...,j$
1b. Select $k-j$ initial seeds from set $X - S$ randomly
2. Each data point $x_i \in S_l$ is assigned to the cluster center $\mu_l$
3. Each data point $x_i \in X - S$ is assigned to its closest cluster center.
4. Each cluster center $\mu_l$ is updated to be the mean of its constituent data points.
5. Repeat 3 and 4 until K-means objective function is optimized.
```

Fig. 4. ISK-means P clustering algorithm

Max-distance searching. In fact, the Seed Set covering j categories has provided a priori knowledge for the rest of the $k-j$ categories. These $k-j$ initial centers should be far away from the j initial centers calculated by equation (2), because the distance between different categories must be as far as possible in a clustering problem. So it is a better idea to choose the farthest data point away from known j initial centers as the $(j+1)$ th initial center than random selection. This calculation method of initial centers is described as below.

- For the j categories covered in $S = \bigcup_{l=1}^{j} S_l, j < k$

$$\mu_l = \frac{1}{|S_l|}\sum_{x \in S_l} x \text{ ’ } l = 1,2,...,j \tag{2}$$

- Choose the farthest data point away from known j initial centers as the $(j+1)$ th initial center:

$$\mu_{j+1} = x_f : when, \max \sum_{l=1}^{j} \left\| x_f - \mu_l \right\| \tag{3}$$

- Repeat step b2 until that all of k initial centers have been obtained.

An example of max-distance searching method is presented in detail in Fig. 5, and we can get a useful conclusion from it. These initial centers generated by max-distance searching have very good dispersion, which is in accord with the most important feature of clustering: distances between different clusters need to be as far as possible.

Furthermore, these initial centers can improve the performance of clustering algorithm. We can call it ISK-means ^{m}, short for the ISK-means algorithm with method of max-distance searching. The algorithm is presented in detail in Fig. 6.

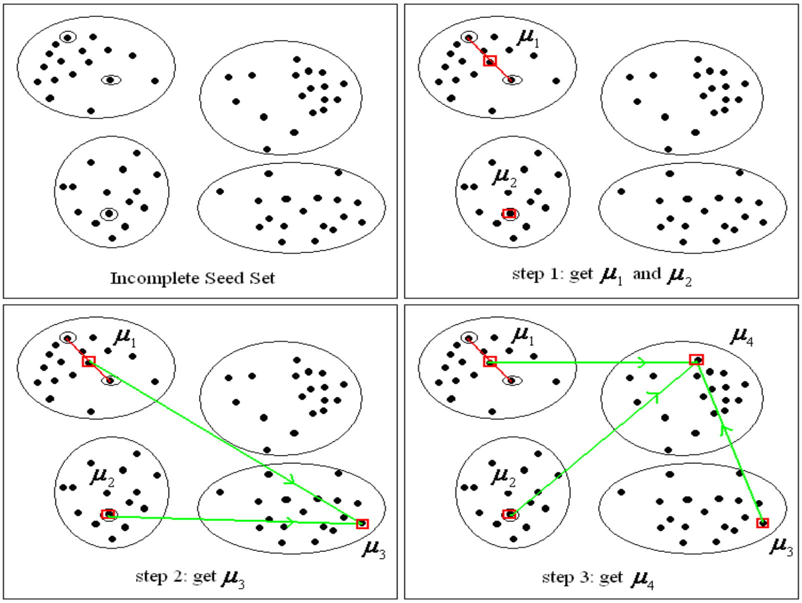
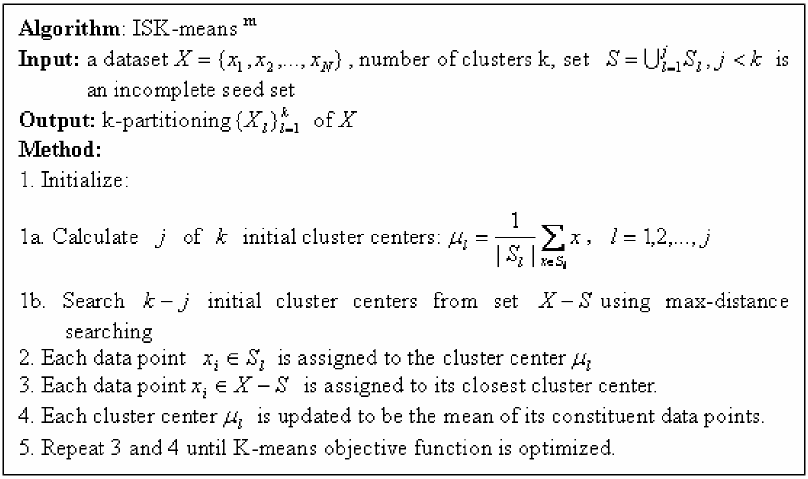


Fig. 5. An Example of Max-distance Searching Method



```
Algorithm: ISK-means $^{m}$
Input: a dataset $X = \{x_1, x_2, ..., x_N\}$, number of clusters k, set $S = \bigcup_{l=1}^{j} S_l, j < k$ is
       an incomplete seed set
Output: k-partitioning $\{X_l\}_{l=1}^{k}$ of $X$
Method:
1. Initialize:
1a. Calculate $j$ of $k$ initial cluster centers: $\mu_l = \frac{1}{|S_l|} \sum_{x \in S_l} x$, $l = 1,2,...,j$
1b. Search $k-j$ initial cluster centers from set $X-S$ using max-distance
    searching
2. Each data point $x_i \in S_l$ is assigned to the cluster center $\mu_l$
3. Each data point $x_i \in X-S$ is assigned to its closest cluster center.
4. Each cluster center $\mu_l$ is updated to be the mean of its constituent data points.
5. Repeat 3 and 4 until K-means objective function is optimized.
```

Fig. 6. ISK-means ^{m} clustering algorithm

4 Experiments

In our experiments, we made use of five UCI datasets [11], i.e. wine, iris, balance, sponge and spectf heart, and one text dataset of practical documents which had been downloaded from the sohu.com website. Its distribution has been shown as table 1. For each dataset, we ran four algorithms: Unsupervised K-means, CSK-means, ISK-means ^{p}, ISK-means ^{m}, and obtained a complete set (for CSK-means) and an incomplete set (for ISK- means ^{p} and ISK-means ^{m}) from the dataset by randomly selection. In all cases, results are averaged over 100 runs, and the accuracy was used to evaluate the performance of these algorithms.

Table 1. Distribution of text dataset

ID	Category Name	Total
1	Finance and Economics	60
2	Sports	80
3	Cars	84
4	Game	86
5	Tourism	76
6	Military affairs	65
7	House Property	45
8	Education	50
9	Healthy diet	66
10	Constellation	38

4.1 Results on UCI Data Sets

Table 2 presents the experimental results on five UCI datasets.

Table 2. Results on UCI datasets

Set	C	I	D	K-means	CSK-means	ISK-Means ^{p}	ISK-means ^{m}
wine	3	178	13	0.59	0.65	0.62	0.63
iris	3	150	4	0.88	0.89	0.88	0.88
balance	3	625	4	0.38	0.45	0.39	0.44
sponge	12	76	45	0.35	0.59	0.38	0.51
spectf	73	267	45	0.25	0.51	0.38	0.46

Note: C-the number of categories, I-the number of instances, D-the number of dimensions.

We can draw the flowing conclusions via analyzing these results:

Conclusion 1: There are few differences among the performances of these four algorithms if the number of clusters is small such as wine, iris and balance datasets. The main reason is that usually unsupervised k-means can get the global optima through numbers of experiments (over 100 times) when the dataset is a convex dataset with small number of clusters. It is difficult to learn more knowledge from a seed set unless the situation that dataset itself has some noise points which are ticklish in unsupervised k-means but noted clearly in the seed set.

Conclusion 2: If the number of clusters is bigger, such as sponge and spectf heart dataset, the performance of CSK-means algorithm exceeds others. It illustrates that given more supervision can promote better performances of algorithms [12-13].

Conclusion 3: Especially when the seed set is incomplete, the performance of ISK-means ^{m} is quite similar with CSK-means whereas higher than ISK-means ^{p}. It shows that the initial cluster centers derived from max-distance search method are good discrete, meanwhile they are nearly as good as initial cluster centers obtained by complete seed set.

4.2 Results on Text Dataset

For the documents dataset, a vocabulary of 7,162 words except the stop words was generated. Each document is represented as a vector in a 7,162 dimensional space, with TFIDF weighting [14].

Table 3 shows the experimental results on the text datasets. We can conclude that CSK-means algorithm can get the global optima in the case of complete seed set, and ISK-means ^{m} algorithm can obtain better results than ISK-means ^{p} in the case of incomplete seed set. However, it is difficult to get a complete seed set actually. In contrast, it is easier to get incomplete one. Therefore, in practical applications it is more valuable to study the algorithms on incomplete seed set basis.

Table 3. Results on text dataset

Set	C	I	D	K-means	CSK-means	ISK-Means ^{p}	ISK-Means ^{m}
text	10	600	7162	0.52	0.81	0.62	0.74

Incomplete seed sets are different in number of partitions and data quantity in per partition. So what effects would different incomplete seed sets have on ISK-means? Based on this question, firstly we changed the number of partitions but maintained seed quantity in per partition as 10%, and we got the experimental results as shown in Fig. 7 by means of operating ISK-means ^{m} and ISK-means ^{p} algorithms again.

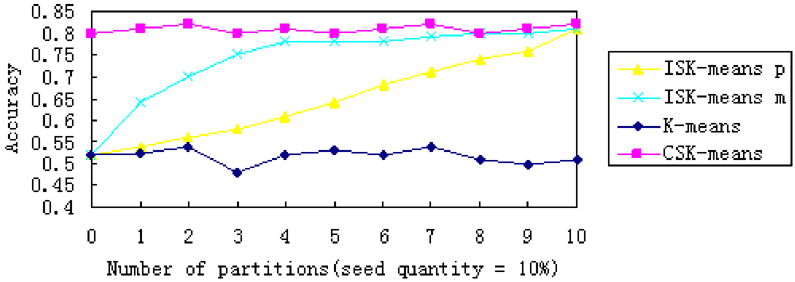


Fig. 7. Incomplete seed sets with different number of partitions

Fig. 7 shows that as the number of partitions increase in complete seed set, the performance of ISK-means ^{p} is a linear growth, whereas for ISK-means ^{m}, the improvement of performance is gradually weak when the number of partitions in incomplete seed set is close to that in complete seed set(actually when the number of partitions > 4). Thus, it can be seen that ISK-means ^{m} algorithm can exploit its advantages when the number of partitions in incomplete seed set is small.

Then we maintained the number of partitions be equal to 4 but changed the seed quantity in per partition, and we got the experimental results as shown in Fig. 8 by means of operating ISK-means ^{m} and ISK-means ^{p} algorithms again.

Fig. 8 shows that when the number of partitions is constant, increasing seed quantity in per partition can also improve performances of algorithms. ISK-means ^{p} can be improved significantly, because it has greatly narrowed down the random searching region with increase of seed quantity. However it is weak to assist ISK-means ^{m} in searching the $k-j$ initial cluster centers by increasing the quantity of seeds in the j seed sets of an incomplete seed set when 10% seeds in per partition have been obtained.

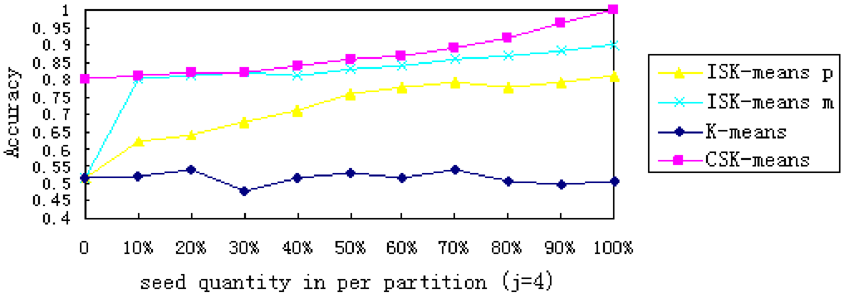


Fig. 8. Incomplete seed sets with the same number of partitions but different number of seeds

5 Conclusion

In this paper, we explained how semi-supervision can be incorporated into the k-means algorithm by means of optimizing initial cluster centers. We divided the supervised information into complete seed sets and incomplete seed sets, so that we could divide semi- supervised k-means into CSK-means and ISK-means. CSK-means can get the initial cluster centers by calculating means of a complete seed set.

However, ISK-means can only get part of initial cluster centers by calculating means of an incomplete seed set. ISK- means ^{m} gets the other centers by max-distance searching method, while ISK-means ^{p} gets them by randomly selecting. Experimental results demonstrated that CSK-means has great performance on complete seed set and ISK-means ^{m} has more advantages than k-means and ISK-means ^{p} on incomplete seed set.

References

1. Olivier, C., Bernhard, S., Alexander, Z.: Semi- Supervised learning, pp. 3–10. MIT Press, Cambridge (2006)
2. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT 1998, Madison, WI, pp. 92–100 (1998)
3. Zhang, T., Ando, R.K.: Analysis of spectral kernel design based semi-supervised learning, pp. 1601–1608. MIT Press, Cambridge (2006)
4. Nizar, G., Michel, C., Nozha, B.: Unsupervised and semi-supervised clustering: a brief survey. In: Proc. of 6th Framework Programme (2005)
5. Basu, S., Bilenko, M., Mooney, R.: A probabilistic framework for semi-supervised clustering. In: Proc. of the 10th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining, pp. 59–68. ACM Press, Seattle (2004)
6. Tao, L., Hongjian, Y.: Semi-supervised learning based on k-means clustering algorithm. Application Research of Computers 27(3), 913–916 (2010)
7. Davidson, I., Basu, S.: Survey of clustering with instance level constraints. ACM Trans. on Knowledge Discovery from Data, 1–44 (2007)
8. Shi, Z.: Semi-supervised model based document clustering: a comparative study. Machine Learning 65(1), 3–29 (2006)
9. Basu, S., Banerjee, A., Mooney, R.J.: Semi- supervised clustering by seeding. In: Proc. of the 19th International Conference on Machine Learning, pp. 19–26 (2002)
10. Wagstaff, K., Cardie, C., Rogers, S.: Constrained k-means clustering with background knowledge. In: Proceedings of the 18th International Conference on Machine Learning, pp. 577–584. Morgan Kaufmann Publishers Inc., San Francisco (2001)
11. Blake, C., Keogh, E., Merz, C.J.: UCI repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine (1998), http://archive.ics.uci.edu/ml/datasets.html
12. Daoqiang, Z., Shiguo, C.: Experimental comparisons of semi-supervised dimensional reduction methods. Journal of Software 22(1), 28–43 (2011)
13. Xiao, Y., Jian, Y.: Semi-supervised clustering based on affinity propagation algorithm. Journal of Software 19(11), 2803–2813 (2008)
14. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)

Fuzzy ID3 Algorithm Based on Generating Hartley Measure

Fachao Li and Dandan Jiang

School of Economics and Management, Hebei University of Science and Technology,
050018, Shijiazhuang, China
lifachao@tsinghua.org.cn, haibian1985@126.com

Abstract. Fuzzy decision tree induction algorithm is an important way with uncertain information. However, the current fuzzy decision tree algorithms do not systematically consider the impact of different fuzzy levels and simply make uncertainty treatment awareness into the selection of extended properties. To avoiding this problem, this paper establishes a generating Hartley measure model based on cut-standard, subsequently, proposes fuzzy ID3 algorithm based on generating Hartley measure model, finally, the results of the experiments indicates that the model is feasible and effective.

Keywords: Generating Hartley measure, the level importance function, generating fuzzy ID3 decision tree algorithm.

1 Introduction

Classification is an important goal and mission of data mining, then the decision tree classification algorithm is an important way to implement classification. It originated in 60 years of the 20th century, was raised by Hunt and others [1] in researching the concept learning system. In 1986, Quinlan [2] proposed the famous decision tree induction algorithm (i. e. ID3 algorithm). The idea of the algorithm is that using information entropy theory based on Hartley measure to select the property with the maximum information gain value as extended attributes in the current sample set, and iterative method is used in growing a new leaf node in the nodes of corresponding subsets, until the non-separable samples, no the remaining property or samples belong to a category. However, this method has some shortcomings, such as tending to choose higher property values, anti-noise ability is poor and can not handle data with missing values and so on. On this basis, then scholars appeared C4.5 [3], SLIQ [4], SPRINT [5] and some scalability decision tree algorithms for the crisp set. However, these algorithms can only deal with data information whose property values and classification values are accurate, can not deal with uncertainty data information about people's thinking and feeling. To avoiding this problem, many scholars have researched some decision tree algorithms in fuzzy environment with fuzzy set theory (e.g. [6-9]), where, Fuzzy ID3 and Min-Ambiguity algorithms are typical. Fuzzy ID3 algorithm uses fuzzy entropy to select extended properties, is an extension of ID3 algorithm, however Min-Ambiguity algorithm uses the uncertainty of possibility distribution to select extended properties, is suitable for fuzzy classification database.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 188–195, 2011.

It is worth noting, the current fuzzy decision tree induction algorithms do not systematically consider the impact of different fuzzy levels, do not simply make uncertainty treatment awareness into the selection of extended properties. To avoiding the lack of current algorithms, generally considering the importance level and size of the cut set, this paper establishes a generating Hartley measure model with structural characteristics in fuzzy sets and uses it in uncertainty measure of fuzzy partition, then proposes extended attribute selection model based on generating Hartley measure model, finally, the results of the experiments indicates that the model is feasible and effective.

In order to facilitate narration, let $\Omega = \{x_1, x_2, \cdots, x_n\}$ be a finite set, $\mathbb{F}(\Omega)$ be the family of all fuzzy subsets on Ω. For $A \in \mathbb{F}(\Omega)$, let $A(x)$ represent the membership function of A, and $A_\lambda = \{x \mid A(x) \geq \lambda\}$ represent $\lambda-$cut of A.

2 Generating Hartley Measure on Fuzzy Subsets

In 1928, Hartley proposed Hartley measure [10], the basic idea is using $H(A) = \log_2 |A|$ to measure the non-prescriptive of the finite set A (where, $|A|$ represents the number of elements in A). In order to establish information treatment method in fuzzy environment, Shafer [11], Higashi and Klir [12] separately established Hartley measure model of fuzzy subsets on different angles (here we make the appointment $\log_2 0 = 0$):

$$H(A) = \log_2 \textstyle\sum_{i=1}^{n} A(x_i)\ , \tag{1}$$

$$H(A) = \textstyle\int_0^1 \log_2 |A_\lambda| \,\mathrm{d}\lambda\ , \tag{2}$$

and discussed the build problems of fuzzy decision tree. Although the two models were both the extension of Hartley measure model, they did not systematically consider the impact of different fuzzy levels.

Because fuzzy set can depict its basic features by level cut set, cut-standard reflects the consistent degree between the elements of cut set and fuzzy concept, so, the reliability of fuzzy decision is different based on the cut set of different levels. Generally, the higher (lower) cut-standard is, the higher (lower) the reliability of fuzzy decision based on the cut set is. In order to reflect systematically the above characteristics of fuzzy information in fuzzy decision, we give the pseudo-elements of fuzzy set and the conception of generating Hartley measure as follows:

Definition 1. Let $W(\lambda)$ be a monotonous level importance weight function on [0,1] (i.e. $W(\lambda)$ is a reflection from [0,1] to $[0, \infty)$, and satisfies: 1) Piecewise continuous, 2) Monotonous nondecreasing, 3) $\int_0^1 W(\lambda)\mathrm{d}\lambda = 1$). For $A \in \mathbb{F}(\Omega)$, we call (**3**) for the pseudo-elements based on the weight function $W(\lambda)$ of A, and call (**4**) for the generating Hartley measure based on the weight function $W(\lambda)$ of A. Where, $|A_\lambda|$ denotes the number of elements in set A_λ, and we make the appointment $\log_2 0 = 0$.

$$M(A) = \textstyle\int_0^1 W(\lambda) |A_\lambda| \,\mathrm{d}\lambda\ , \tag{3}$$

$$H(A)=\int_0^1 W(\lambda)\log_2 | A_\lambda | \mathrm{d}\lambda, \quad (4)$$

It is easily to see: 1) When A is the crisp set, $M(A)=| A |$, $H(A)=\log_2 | A |$; 2) When $W(\lambda)\equiv 1$, $M(A)=\sum_{i=1}^{n} A(x_i)$, (4) is namely (2); 3) When $A\subset B$, $M(A)\leq M(B)$, $H(A)\leq H(B)$.

Theorem 1. Let $\Omega=\{x_1, x_2,\cdots,x_n\}$, $A\in \mathbb{F}(\Omega)$, and $A(x_i)>0, i=1,2,\cdots,n$, Ω_1, $\Omega_2,\cdots,\Omega_m$ be a partition of Ω which satisfies the following conditions: 1) For each Ω_k, when $x, y\in \Omega_k$, $A(x)=A(y)\triangleq \lambda_k$; 2) For any different Ω_i and Ω_j, when $x\in \Omega_i$, $y\in \Omega_j$, $A(x)\neq A(y)$; 3) When $\lambda_1<\lambda_2<\cdots<\lambda_m$, for any level importance weight function $W(\lambda)$, then

$$M(A)=\sum_{j=1}^{m} w_j (n-\sum_{k=1}^{j-1} n_k), \quad (5)$$

$$H(A)=\sum_{j=1}^{m} w_j \log_2 (n-\sum_{k=1}^{j-1} n_k). \quad (6)$$

Where, $w_j=\int_{\lambda_{j-1}}^{\lambda_j} W(\lambda)\mathrm{d}\lambda$, $n_j=| \Omega_j |$ represents the number of elements in Ω_j, here we make the appointment $\lambda_0=0$.

This theorem can be directly proved by the properties of integration. It gives a method of calculating generating Hartley measure value, when calculating $H(A)$, let Ω be $\mathrm{supp}A=\{x | x\in \Omega, A(x)>0\}$.

Intuitively speaking, $W(\lambda)$ depicts the impact degree of cut-standard λ to the research question. In reality, according to the properties of the research questions and different emphases, we can select the specific form.

3 The Standardized Description of Fuzzy ID3 Algorithm

In the recursive and structure process of decision tree, fuzzy ID3 algorithm [13] adopts the divide and conquer strategy, i.e. using the fuzzy classification entropy of the attribute-value as heuristic function of expanded attributes, selecting the smallest classification fuzzy entropy of the attribute as expanded attributes. Here, we make the appointment: $E=\{1, 2, \cdots, N\}$ denotes the training set, $A=\{A_1, A_2,\cdots, A_m\}$ represents condition attribute set, range $(A_i)=\{A_{i1}, A_{i2},\cdots, A_{im_i}\}$ represents the range of the attribute A_i, range $(C)=\{c_1, c_2,\cdots, c_m\}$represents the range of the decision attribute, then: 1) In fuzzy environment, attribute-value can be considered as fuzzy set to the training set $E=\{1, 2, \cdots, N\}$ (i.e. N dimensional fuzzy vector on E); 2) The nodes of the decision tree can be considered as fuzzy set on E. According to the above analysis and appointment, the following is the implementation process of fuzzy ID3 algorithm (where, for fuzzy set B on E, $M^*(B)=\sum_{i=1}^{N} B(i)$ represents the cardinality of B):

Step 1. For the current node D: a) For the given leaf criterion $\delta(0<\delta<1)$, if there is a $k'\in \{1, 2,\cdots, n\}$ such that $f_{k'}(D)\geq \delta$, then D is a leaf node; b) If father node of D has used all the condition attributes in branching process, then D is a leaf node. Where, $f_k(D)=M^*(D\cap c_k)/M^*(D)$.

Step 2. If D is not leaf node, then calculate the fuzzy classification entropies $FE(D, A_i)$ of the condition attributes A_i which are not used by father node of D in branching process, and select the smallest classification fuzzy entropy to the attribute as expanded attributes(i.e. according to this attribute to branch).

$$FE(D, A_i) = -\sum_{j=1}^{k_i} \frac{m_{ij}}{m_i} E(D \cap A_{ij}) \,, \tag{7}$$

$$E(D \cap A_{ij}) = -\sum_{k=1}^{n} \frac{m_{ijk}}{\bar{m}_{ij}} \log_2 \frac{m_{ijk}}{\bar{m}_{ij}} \,, \tag{8}$$

Where, $m_{ij} = M^*(D \cap A_{ij})$, $m_i = \sum_{j=1}^{k_i} m_{ij}$, $m_{ijk} = M^*(D \cap A_{ij} \cap c_k)$, $\bar{m}_{ij} = \sum_{k=1}^{n} m_{ijk}$.

Step 3. Repeat step 1 and step 2, until did not branch.

Step 4. Conversion the generating tree to the corresponding rules.

4 Generating Fuzzy ID3 Algorithm

From the implementation process of fuzzy ID3 algorithm, we can see this algorithm is the extension of classical ID3 algorithm, fuzzy classification entropy is the extension of crisp classification entropy. But because fuzzy classification is non-additivity, then modify (8) to

$$E(D \cap A_{ij}) = -\sum_{k=1}^{n} \frac{m_{ijk}}{\bar{m}_{ij}} \log_2 \frac{m_{ijk}}{m_{ij}} \,. \tag{8'}$$

Obviously, this formula more satisfies the objective situation, and (7) is still the extension of crisp classification entropy. But it is worth noting that whether (8) or (8'), (7) does not reflect the impact of different membership degree. With the discussion of Part 2, in order to reflect the impact of different membership degree, we can use (3) as the cardinality measure model of fuzzy set, use (4) as the non-prescriptive measure model of fuzzy set, when adjust m_{ij} in (7) to (**7'**), adjust $E(D \cap A_{ij})$ in (**8'**) to (**8''**).

$$m_{ij} = M(D \cap A_{ij}) = \int_0^1 W(\lambda) \,|\, (D \cap A_{ij})_\lambda \,|\, \mathrm{d}\lambda \,, \tag{7'}$$

$$E(D \cap A_{ij}) = -\sum_{k=1}^{n} \frac{m_{ijk}}{\bar{m}_i} [H(D \cap A_{ij} \cap c_k) - H(D \cap A_{ij})] \,. \tag{8''}$$

Where, $m_{ijk} = M(D \cap A_{ij} \cap c_k) = \int_0^1 W(\lambda) \,|\, (D \cap A_{ij} \cap c_k)_\lambda \,|\, \mathrm{d}\lambda$, $\bar{m}_{ij} = \sum_{k=1}^{n} m_{ijk}$, $H(B) = \int_0^1 W(\lambda) \log_2 |\, B_\lambda \,|\, \mathrm{d}\lambda$.

Obviously, after the adjustment, according to fuzzy decision tree algorithm which is formed from step 1 to step 4 in Part 3 (i.e. generating fuzzy ID3 algorithm), this algorithm is still the extension of classical ID3 algorithm, we can adjust the weight function $W(\lambda)$ to reflect the importance of different membership degree.

5 Analysis and Comparison of Experimental Results

In this section, with a fuzzy attribute-value learning problems from examples (Table 1), we will compare and analyze the features and performance of generating fuzzy ID3 algorithm, where, the fuzzy condition attributes are Temperature, Outlook, Humidity and Wind, the fuzzy decision attribute is Class. Figure 1 is the classification decision tree based on fuzzy ID3 algorithm [14], while the classification decision trees based on generating fuzzy ID3 algorithm are Figure 2~4, where, the leaf-standard δ is 0.8.

Table 1. Training Set with Fuzzy Representation

	Temperature			Outlook			Humidity		Wind		Class		
	Hot	Mild	Cool	Suny	Cloudy	Rain	Humid	Normal	Windy	Not-windy	Volley-ball	Swimming	Weight-lifting
1	0.7	0.2	0.1	1.0	0.0	0.0	0.7	0.3	0.4	0.6	0.0	0.6	0.4
2	0.6	0.2	0.2	0.6	0.4	0.0	0.6	0.4	0.9	0.1	0.7	0.6	0.0
3	0.0	0.7	0.3	0.8	0.2	0.0	0.2	0.8	0.2	0.8	0.3	0.6	0.1
4	0.2	0.7	0.1	0.3	0.7	0.0	0.8	0.2	0.3	0.7	0.9	0.1	0.0
5	0.0	0.1	0.9	0.7	0.3	0.0	0.5	0.5	0.5	0.5	1.0	0.0	0.0
6	0.0	0.7	0.3	0.0	0.3	0.7	0.3	0.7	0.4	0.6	0.2	0.2	0.6
7	0.0	0.3	0.7	0.0	0.0	1.0	0.8	0.2	0.1	0.9	0.0	0.0	1.0
8	0.0	1.0	0.0	0.0	0.9	0.1	0.1	0.9	0.0	1.0	0.3	0.0	0.7
9	1.0	0.0	0.0	1.0	0.0	0.0	0.4	0.6	0.4	0.6	0.4	0.7	0.0
10	0.7	0.2	0.1	0.0	0.3	0.7	0.8	0.2	0.9	0.1	0.0	0.3	0.7
11	0.6	0.3	0.1	1.0	0.0	0.0	0.7	0.3	0.2	0.8	0.4	0.7	0.0
12	0.2	0.6	0.2	0.0	1.0	0.0	0.7	0.3	0.7	0.3	0.7	0.2	0.1
13	0.7	0.3	0.0	0.0	0.9	0.1	0.1	0.9	0.0	1.0	0.0	0.4	0.6
14	0.1	0.6	0.3	0.0	0.9	0.1	0.7	0.3	0.7	0.3	1.0	0.0	0.0
15	0.0	0.0	1.0	0.0	0.3	0.7	0.2	0.8	0.8	0.2	0.4	0.0	0.6
16	1.0	0.0	0.0	0.5	0.5	0.0	1.0	0.0	1.0	0.0	0.7	0.6	0.0

The related information (including the number of node, the number of leaf, depth and average accuracy) from the above decision tree is listed in the Table 2. We can know from the table, the value of the decision tree based on generating fuzzy ID3 algorithm is better than the value of the decision tree based on fuzzy ID3 algorithm. Where, for the three selected weight function $W(\lambda)$, when $W_2(\lambda)=1.5\sqrt{\lambda}$, the category superiority of decision tree is the strongest. This shows, when selecting appropriate weight function $W(\lambda)$, that can generate the decision tree with more effective classification and higher accuracy. Thus indicates that the generating fuzzy ID3 algorithm is feasible and effective.

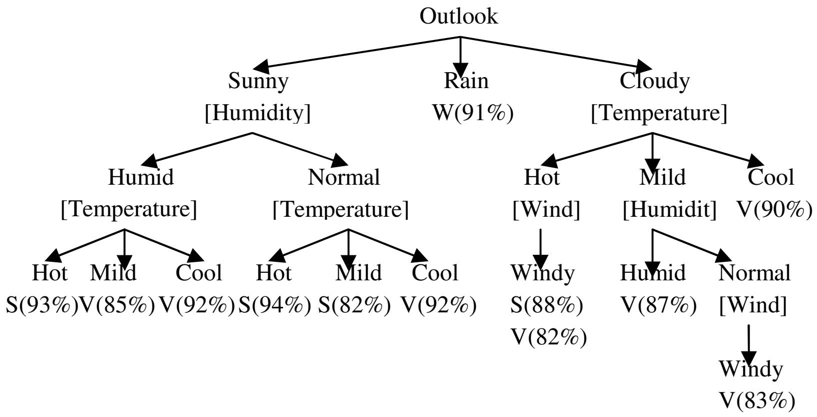


Fig. 1. The decision tree based on fuzzy ID3 algorithm

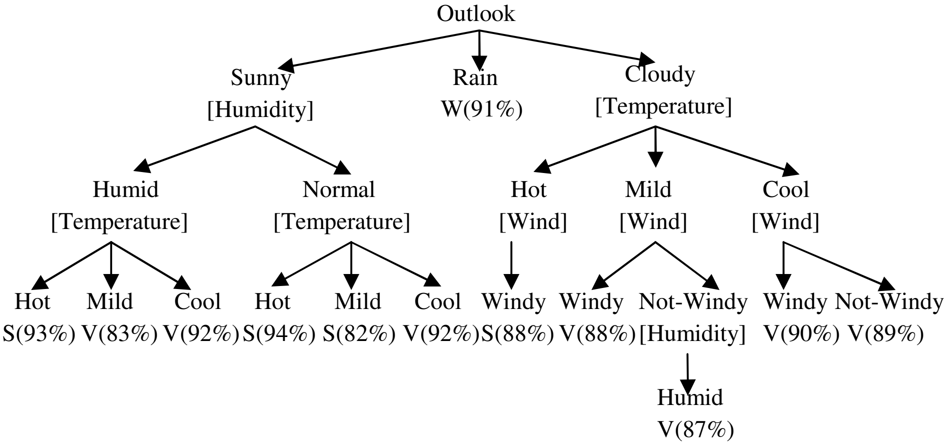


Fig. 2. The decision tree based on generating fuzzy ID3 algorithm for $W_1(\lambda) = 2\lambda$

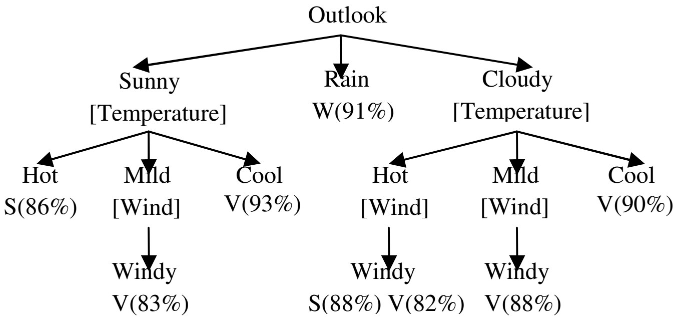


Fig. 3. The decision tree based on generating fuzzy ID3 algorithm for $W_2(\lambda) = 1.5\sqrt{\lambda}$

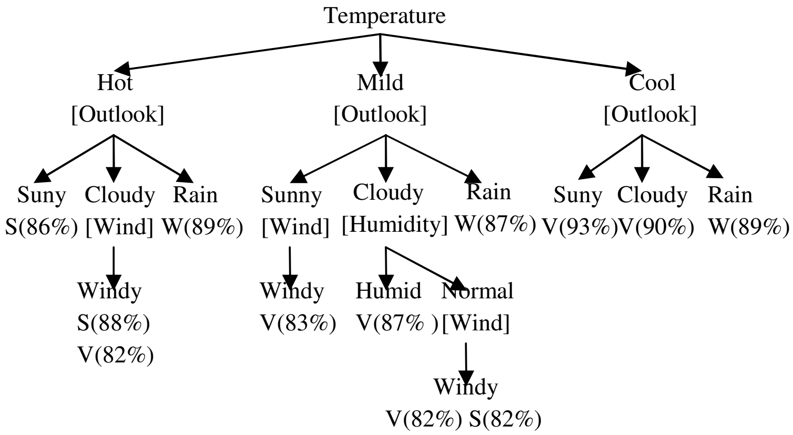


Fig. 4. The decision tree based on generating fuzzy ID3 algorithm for $W_3(\lambda) = 3\lambda^2$

Table 2. Database Summary

Database	Fig. 1	Fig. 2	Fig. 3	Fig. 4
The number of node	7	7	6	6
The number of leaf	11	11	7	10
Depth	5	5	4	5
Average accuracy(%)	88.25	89.10	87.63	86.45

6 Conclusions

In this paper, with the discussion about level importance and size of the cut set, we consider the different impact of different fuzzy levels and simply make uncertainty treatment awareness into the selection of extended properties. First, we establish a generating Hartley measure model in fuzzy sets and use it in uncertainty measure of fuzzy partition, then propose the extended attribute selection model based on generating Hartley measure model, i.e. generating fuzzy ID3 algorithm. Finally, the results of the experiments indicates that we can generate the decision tree with more effective classification and higher accuracy by using this algorithm. Thus illustrates the model is feasible and effective, preliminarily implements the optimization of fuzzy ID3 algorithm. Of course, about a more detailed discussion and more extensive example shows of this model, we will continue to work in the future.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (71071049, 70871036) and the Natural Science Foundation of Hebei Province (F2011208056).

References

1. Davis, L.D.: Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York (1991)
2. Quinlan, J.R.: Induction of decision trees. Machine Learning, 81–106 (1986)
3. Quinlan, J.R.: C4.5: Programs for Machine Learning, pp. 1–131. Morgan Kaufman, San Francisco (1993)
4. Metha, M., Agrawal, R., Rissanen, J.: SLIQ: A Fast Scalable for Classifier for Data Mining. In: Apers, P.M.G., Bouzeghoub, M., Gardarin, G. (eds.) EDBT 1996. LNCS, vol. 1057, pp. 18–32. Springer, Heidelberg (1996)
5. Shafer, J., Agrawal, R., Mehta, M.: SPRINT: A Scalable Parallel Classifier for Data Mining. In: Proceedings of the 22nd International Conference on Very Large Databases, pp. 544–555. Morgan Kauffman, San Mateo (1996)
6. Yuan, Y., Shaw, M.J.: Induction of fuzzy decision trees. Fuzzy Sets and System 69, 125–139 (1995)
7. Wang, X.G., Yeung, D.S., Tsang, E.C.C.: A comparative study on heuristic algorithms for generating fuzzy decision trees. IEEE Transactions on Systems. Man and Cybernet. Ics-part B: Cybernetics 31(2) (April 2001)
8. Zhang, X., Zhao, H.: Fuzzy decision fusion algorithm of uncertain information. Journal of Northeastern University 7(25), 657–660 (2004)
9. Zhai, J., Wang, H.: A fuzzy classification algorithm based on fuzzy entropy. Computer Engineering and Applications 20(46), 176–180 (2010)
10. Eyrvl, H.: Transmission of Information. Bell Systems Technical Journal 17, 535–563 (1928)
11. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
12. Higashi, M., Klir, G.J.: Measures of uncertainty and information based on possibility distributions. Internat. J. Gen. Systems 9, 43–58 (1983)
13. Umanol, M., Okamoto, H., Hatono, I., Tamura, H., Kawachi, F., Umedzu, S., Kinoshita, J.: Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems. In: IEEE Int. Conf. on Fuzzy Systems, June 26–29, pp. 2113–2118 (1994)
14. Zhai, J., Wang, X., Zhang, S.: The fusion of multi fuzzy decision tree based on fuzzy Integral. Computer Research and Development 3(46), 470–477 (2009)

A Technique for Improving the Performance of Naive Bayes Text Classification

Yuqian Jiang, Huaizhong Lin*, Xuesong Wang, and Dongming Lu

College of Computer Science and Technology, Zhejiang University,
Zheda Road 38, Hangzhou, Zhejiang, China
{lanseyu_19,iamsnowpeer}@163.com,
{linhz,ldm}@zju.edu.cn

Abstract. Naive Bayes classifier is widely used in text classification tasks, and it can perform surprisingly well, it is often regarded as a baseline. But previous researches show that the skewed distribution of training collection may cause poor results in text classification. This paper presents a new method to deal with this situation. We introduce a conditional probability which takes into account both the information of the whole corpus and each category. Our proposed method performs well in the standard benchmark collections, competing with the state-of-the-art text classifiers especially for the skewed data.

Keywords: Text classification; naive Bayes classifier; skewed data; conditional probability.

1 Introduction

Text classification is a task which assigns the text documents with predefined categories, and it's widely used in natural language processing. Naive Bayes is based on an independence assumption that the features in a document are independent. Although this assumption is violated in natural language, Domingos et al.[2] show that the naive bayes classification can still perform surprisingly well.

There are mainly two models to model the generation of documents: multivariate Bernoulli model and multinomial model. Multivariate Bernoulli model uses a binary attribute to indicate whether the special word occurs in the document[5]; while multinomial model incorporates the frequency information compared with multivariate Bernoulli model[9]. Recent researchers usually regard the multinomial model as a baseline in text classification.

Although the naive Bayes is frequently used in text classification, there exist a lot of problems to improve. One systemic problem is that the training sample may be not uniform, some researchers call this occasion as skewed data[4], when one class has more training examples than another, or the training documents in one class are longer than another, naive Bayes usually prefers one class over the other[9]. Another problem is that the naive Bayes assumption ignores the relationship between words, meanwhile the parameter estimation is too rough.

* Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 196–203, 2011.

For skewed data, Sang-Bum Kim et al. [4] proposed a per-document length normalization approach based on multivariate Poisson model to improve the influence by long documents and short documents. Jason D. M. Rennie et al.[9] introduced a "complement class" to estimate parameters using data from all classes except C_j. For parameter estimation, Kamal Nigam et al.[7] used maximum entropy to estimate probability distributions. katz et al.[3] found that occurrences of the same word in a document are revelent, they called this occasion burstiness. GuoQiang[8] proposed a transformation to the word frequency, which aimed to push down the word frequency to some extent. Ben Allison[1] advocated the use of a joint beta-binomial distribution for word counts in documents for Classification. Meanwhile feature selection is an important preprocessing step for dimension reduction. A statistical methods CHIR was proposed to select features[6]. Karl-Michael Schneider proposed a feature scoring function named CRQ(Cluster Representation Quality) to select features, this function is based on distribution clustering[11]. But with the feature selection method, one should determine the number of feature selected for each class, while how to determine the optimal number is still a challenge.

In this paper we propose a method to improve naive Bayes classification by introducing a new conditional probability based on the Bayes' theorem. We use this probability to estimate the probability of a particular feature belonging to a given class, and its estimation takes into account the global and local information of the training data. The experiments show that our proposed method performs well, especially for the skewed data.

This paper proceeds as follows. Section 2 presents the general framework for standard naive Bayes classifier. Then our proposed approach is discussed in Section 3. Experimental results are presented in Section 4. Finally, Section 5 concludes our paper and gives a future plan.

2 The Naive Bayes Probability Model

In the Naive Bayes probabilistic model, we define C as the set of predefined categories, C consists of m components, $C=\{c_1,c_2,\ldots,c_m\}$, we define $P(c_j|d_i)$ as the probability that a document d_i belongs to a class c_j, the classifier selects the class with the maximum probability as the result class, $P(c_j|d_i)$ is calculated by the Bayes' theorem as follows:

$$P(c_j|d_i) = \frac{P(c_j)P(d_i|c_j)}{P(d_i)} \tag{1}$$

where $p(d_i|c_j)$ represents the distribution of documents in each class, it cannot be estimated directly. But based on "naive Bayes assumption", the document d_i is treated as a multinomial distribution of features w_k with the length of d_i[5], thus we estimate $p(d_i|c_j)$ as follows:

$$p(d_i|c_j) = \prod_{k=1}^{n} P(w_k|c_j)^{N_{ki}} \quad \textit{and} \quad \sum_{j=1}^{m} P(w_k|c_j) = 1 \tag{2}$$

Let N_{ki} be the counts of w_k in document d_i, m is the number of categories, n is the vocabulary size of document d_i. Since d_i is constant and can be ignored:

$$P(c_j|d_i) = P(c_j) \prod_{k=1}^{n} P(w_k|c_j)^{N_{ki}} \tag{3}$$

Where $P(c_j)$ and $P(w_k|c_j)$ are priori probabilities estimated from the training data. $P(c_j)$ represents the probability of category c_j, $P(w_k|c_j)$ represents the probability of the feature w_k in category c_j, it can be calculated as follows:

$$P(w_k|c_j) = \frac{N_{kj} + 1}{N_j + m} \tag{4}$$

Where N_{kj} is the counts of occurrence for w_k in category c_j, N_j is the numbers of words for category c_j.

We can see from formula (4), the multinomial model treats all the training documents for a given class as a big document, and discards the affection by the length of documents and the number of training documents in each category.

3 Proposed Method for Improving Naive Bayes Text Classifier

In standard naive Bayes classifier, we mainly calculate the conditional probability $P(w_k|c_j)$. While in our approach we introduce a new conditional probability $P(c_j|w_k)$ by using the Bayes' theorem again:

$$P(w_k|c_j) = \frac{P(c_j|w_k)P(w_k)}{P(c_j)} \tag{5}$$

Since d_i is constant and can be ignored, (2) is rewritten using (5) as:

$$P(c_j|d_i) = P(c_j) \prod_{k=1}^{n} \left(\frac{P(c_j|w_k)}{P(c_j)} \right)^{N_{ki}} \tag{6}$$

In traditional naive Bayes classification, we mainly estimate the probability of a given class generating a word, this estimation is under the condition of a given class. We can say it is estimated in a local view without considering the whole training data, even in a local view, it still does not take into account the distribution in each category. We know a short document usually consists of more essential features to express its topic, and a long document usually has more unnecessary words, it decreases the probability of category c_j with more long documents generating a feature w_k. Our proposed method transforms to estimate the probability of a given feature belonging to a category, which can be calculated both in a global view and a local view, we will explain it later.

In (6), $P(c_j|w_k)$ reflects the relationship between features and categories , that is how likely a particular feature given belongs to a category, it is somewhat similar to the concept of dominance which is used in a temporal weighting function mentioned by L. Rocha et al.[10]. The normal way to calculate $P(c_j|w_k)$ is:

$$P(c_j|w_k) = \frac{d_{kj}+1}{D_k+m} \tag{7}$$

Let d_{kj} be the number of the documents in category c_j with the feature, and D_k be the total number of the documents in all categories with the feature. We can see that this calculation does not take into account the number of documents in each category, if a category c_i has more documents than another category c_j, the probability for w_k in c_i is likely higher than in c_j, even though w_k is more likely to be for c_j. We list two other methods to calculate $P(c_j|w_k)$ as follows:

$$P_1(c_j|w_k) = \frac{n_{kj}+1}{N_c+m} \tag{8}$$

$$P_2(c_j|w_k) = \frac{n_{kj}+1}{N_k+m} \tag{9}$$

Where n_{kj} represents the total counts of occurrence for w_k in category c_j, N_c and N_k represent the number of words in all categories and the total counts of occurrence for w_k in all categories. In order to smooth the effect of long documents, we transform (9) to (10) with N_{cj} which represents the numbers of words for category c_j :

$$P_2(c_j|w_k) = \frac{n_{kj}}{(N_k+m)*N_{cj}} \tag{10}$$

Let's analyze the above two formulas. Formula (8) treats the whole training data as an entity, and calculates the proportion of the counts of occurrence for w_k in category c_j to all words in the whole training data, we call it as global information here, Formula (10) is mainly affected by the numbers of words in the category, and we call it as local information here. Then we estimate $P(c_j|w_k)$ using both global information and local information as follows:

$$P(c_j|w_k) = P_1(c_j|w_k)^{\alpha} P_2(c_j|w_k)^{1-\alpha} \tag{11}$$

In (11), since the two parts have different intention in estimating $P(c_j|w_k)$, we use a parameter α to balance the effect of two functions, and α is determined empirically, we will discuss the relationship between the distribution of corpus and different values of α in the following experiments. Using (11) in (6) is:

$$P(c_j|d_i) = P(c_j)\prod_{k=1}^{n}\left(\frac{P_1(c_j|w_k)^{\alpha}P_2(c_j|w_k)^{1-\alpha}}{P(c_j)}\right)^{N_{ki}} \tag{12}$$

4 Experimental Results and Analysis

4.1 Data and Evaluation Measure

We run experiments on two commonly used Chinese corpus in text categorization, one is supplied by Shanghai fudan University and the other is by Sogou

Lab. The first collection consists of ten categories, it is separated to have 1892 training documents, and 934 test documents, each training category has obviously different number of documents, it is that the distribution is skewed. The second collection consists of nine categories, 991 messages from each category were chosen randomly as training set, and 7200 for testing, the second collection is uniform relatively to the first . And we use F1 measure of each category to measure the classification performance.

4.2 Experimental Results and Discussion

We compare our classification performance with standard naive Bayes and a per-document length normalization approach proposed by Sang-Bum Kim et al.[4]. This approach normalizes the term frequencies in each document according to the document length, here we choose one simple and effective length normalization method for the experiment:

$$N^1_{ki} = \frac{N_{ki}}{NF} \tag{13}$$

In (13) NF indicates a normalization factor

$$NF = \beta * avdl + (1 - \beta) * dl_i \tag{14}$$

Where N_{ki} indicates the number of word w_k occurrence in document d_i, $avdl$ and dl_i represent the average number of words in a document d_i and the number of words in d_i. Here we make a small change for convenience. We use the average number of words and the number of words in a category instead of in a document, and the number of word w_k occurrence in a category instead of in a document.

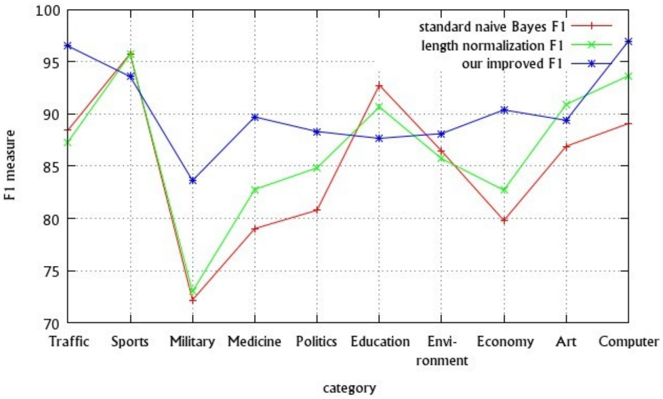


Fig. 1. Comparison of the performance of different classifications on collection from Fudan University

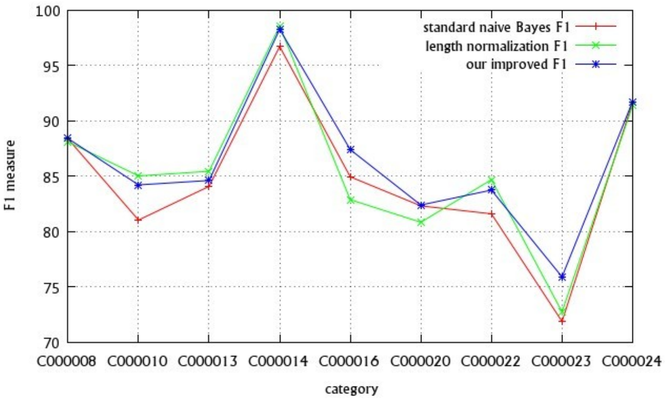


Fig. 2. Comparison of the performance of different classifications on collection from Sogou Lab

In Figure 1, we observe that our proposed classifier significantly outperforms the traditional multinomial naive Bayes and the length normalization method at least when this collection is used. The highest improvement achieves to 11%. In Figure 2, our method still outperforms the traditional multinomial naive Bayes about 2% improvements, but not as significantly as the first experiment. Considering the data sets, the distribution of the second collection is uniform. Thus from Figure 1 and Figure 2, it is possible to conclude that our method is effective, especially for the skewed data set.

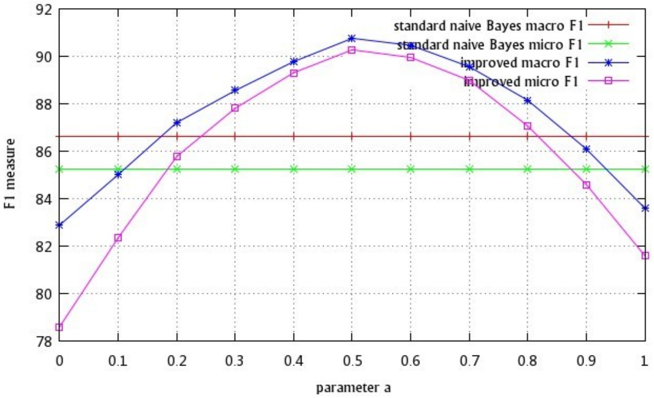


Fig. 3. Performance with various values of α on collection from Fudan University

We continue to discuss the relationship between the parameter α and the distribute of collections. Figure 3 and Figure 4 show the performance of our method with various values of parameter α. Figure 3 shows that the micro F1

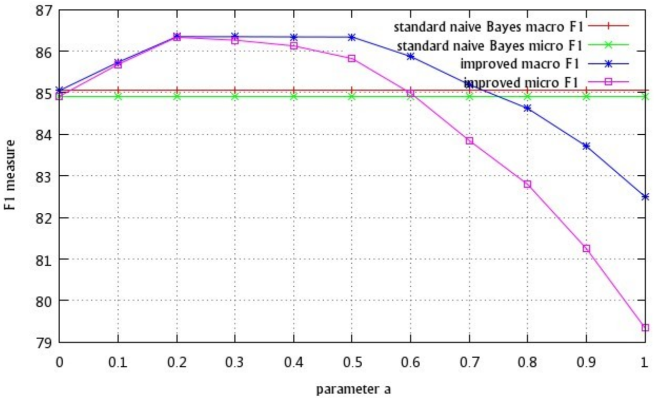


Fig. 4. Performance with various values of α on collection from Sogou Lab

and macro F1 at α of 0.5 achieves the best performance for the first collection, and the best range of α is between 0.2 and 0.9. Figure 4 shows the micro F1 and macro F1 at α of 0.3 achieves the best performance for the second, and the best range of α is between 0 and 0.6.

From the experiments we can see that when the training data is skewed, a high value α means that is $P(c_j|w_k)$ is mainly determined by the global information, our explanation for this conclusion is that when each category has different number of training documents, the information each category providing is different obviously, thus treating the whole data as a entity is more suitable than each category, and when the number of each category is the same, the probability $P(c_j|w_k)$ is mainly determined by the local information.

5 Conclusion

Most of the recent text classification researches focus on addressing problems such as feature selection, length normalization and dimensionality reduction. In this paper, we propose a approach with a novel conditional probability by using the Bayes' theorems twice, our method combines the global information and local information to smooth the influence brought by the distribution of training data. Experiments on two collections show that our proposed approach can significantly improve the performance of classification, especially for the skewed data, and the importance of global and local information is related to the distribution of training data.

Many researches of further work remain. Our experiments results indicate that there is no remarkable effect when the distribution of collection is uniform. We will develop more suitable method to estimate the novel conditional probability, and break the limit for uniform distributed data set.

Acknowledgments. This research has been supported by the National Science and Technology Support Programs(no:2009BAH43B02, no:2009BAH43B05 and no:2010BAK67B08) and the Public Technology Research Projects in Zhejiang Province(no:2010C33151).

References

1. Allison, B.: An improved hierarchical Bayesian model of language for document classification. In: Proceedings of the 22nd International Conference on Computational Linguistics-, vol. 1, pp. 25–32. Association for Computational Linguistics (2008)
2. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29(2), 103–130 (1997)
3. Katz, S.: Distribution of content words and phrases in text and language modelling. Natural Language Engineering 2(1), 15–59 (1996)
4. Kim, S., Han, K., Rim, H., Myaeng, S.: Some effective techniques for naive bayes text classification. IEEE Transactions on Knowledge and Data Engineering, 1457–1466 (2006)
5. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. In: AAAI-1998 Workshop on Learning for Text Categorization, Citeseer, vol. 752, pp. 41–48 (1998)
6. Meena, M., Chandran, K.: Naive Bayes text classification with positive features selected by statistical method. In: First International Conference on Advanced Computing, ICAC 2009, pp. 28–33. IEEE, Los Alamitos (2009)
7. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: IJCAI 1999 Workshop on Machine Learning for Information Filtering, Citeseer, vol. 1, pp. 61–67 (1999)
8. Qiang, G.: An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification. In: 2010 Second International Conference on Computer Research and Development, pp. 699–701. IEEE, Los Alamitos (2010)
9. Rennie, J., Shih, L., Teevan, J., Karger, D.: Tackling the poor assumptions of naive bayes text classifiers. In: Machine Learning-International Workshop Then Conference-, vol. 20, p. 616 (2003)
10. Rocha, L., Mourão, F., Pereira, A., Gonçalves, M., Meira Jr., W.: Exploiting temporal contexts in text classification. In: Proceeding of the 17th ACM Conference on Information and Knowledge Management, pp. 243–252. ACM, New York (2008)
11. Schneider, K.: Techniques for improving the performance of naive Bayes for text classification. In: Computational Linguistics and Intelligent Text Processing, pp. 682–693 (2005)

Mapping Data Classification Based on Modified Fuzzy Statistical Analysis

Yi Cheng, Mingxia Xie, and Jianzhong Guo

Zhengzhou Institute of Surveying and Mapping, Zhengzhou, China
chxycy@126.com

Abstract. Through analyzing and researching traditional mapping data classification , and considering the fuzzy of classification, then a modified mapping data classification is put forward, which is based on fuzzy set. First, gain fuzzy sample set by expert system, and compute sample distribution function by statistical analysis. Then, utilize distribution function to gained fuzzy membership function, and working out the fuzziest point by membership function. Finally, in according to the most fuzzy point, and achieve to classify mapping data. The proposed method solves the problem of misconstruction of the original data, not being generally used, complexity of computing and fuzzy classification in the traditional mapping data classification method and practical process.

Keywords: fuzzy set, fuzzy membership function, the fuzziest point, mapping data classification, Statistical Analysis, P-P probability map.

1 Introduction

Both classical formal logic and modern mathematical logic are accurate binary logics, which are applying for precision objects and phenomena research. If they are used to research fuzzy objects, a simplify operation are needed to change the fuzzy objects to precision ones. Some fuzzy attributes are abandoned in the operation and a manmade boundary line is designed to separate the object. However, the artificial separate line is a kind of distorts to objects and this distort will become seriously near the line especially. So a conclusion is made clearly that binary logics are not adapted to complex objects. Based on fuzzy set, a modified mapping data classification is put forward in the paper to avoid distort by hard-separate.

2 Basic Definitions

2.1 Mapping Data Classification Theory

Actually, classification is the common method to generalize the characteristics of objects. Based on the classification of spatial and attribute data, thematic maps are designed to reveal the regulations of objects distribution and development.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 204–210, 2011.

(1) Mapping Data Classification Fundamental

a. Get the number of classification

- In order to fix the number of classification, the map use and the requirements of estimate accuracy should be taken into consideration.
- In order to fix the number of classification, the emphasis of area distributing character should be taken into consideration.
- In order to fix the number of classification, the influence of map scale should be taken into consideration.
- In order to fix the number of classification, the distributing character of data should be keep well.
- On condition of the map use requirements of statistical accuracy, the number of classification should be decreased. And
- The method of data expression should be taken into consideration also.

b. Get the boundary

- The fixed boundary must meet the difference in the same class is minimum and the difference between classes is maximum.
- There is at least one in any class and any data belong to a class.
- The boundary might be contiguous or non-contiguous decide by the data character.
- It is better to select a contiguous boundary to help user read and memory. And
- Boundary should be nearest integer to enhance the readability.

(2) The Traditional Methods of Mapping Data Classification

The traditional methods of mapping data classification include sequence classification, series classification, mean-standard deviation classification, nested mean classification, quintile classification and stepwise pattern recognition classification. All of them will lead to hard plot to data and it is full of fuzzy in real classification problems and the detailed algorithms have been proposed in [2].

2.2 Fuzzy Set Theory

If $\widetilde{A}$ is mapping of discourse domain X to $[0,1]$, $\widetilde{A}: X \to [0,1], x \mapsto \widetilde{A}(x)$, that $\widetilde{A}$ is a fuzzy set of X and $\widetilde{A}(x)$ is member function of $\widetilde{A}$. As we know, $\widetilde{A}$-fuzzy set of X is a function of $X \to [0,1]$, so that the first task is to define the member function $\widetilde{A}(x)$. We can define $\widetilde{A}(x)$ by the methods of fuzzy statistic, dualism-based comparing sort, set-valued statistics, interpolation and increment method, etc.

3 The Method of Mapping Data Classification Based on Modified Fuzzy Statistical Analysis Selecting a Template

3.1 Problems Existed in the Traditional Mapping Data Classification

Both series classification and progression classification are simple to calculate and the change of boundary is regularly. They are suitable for understand and comparative analysis. However, boundary is separated from data distributing character so that information of raw data might be distorted in these ways. Average value-standard deviation is only suitable for normal distribution or approximating normal distribution. Nested average value is suitable for average distribution and the number of class must be even number. Quintile distribution classification is suitable for the classification of level data. Maintaining the Integrity of the Specifications.

3.2 The Traditional Multivariate Statistical Analysis Models Used in Classification [3]

We set the research object as y and the independent variables related with y are

$$x_i = (x_{1i}, x_{2i}, \cdots, x_{mi}), i = 1, 2, \cdots, n,$$

where n is the number of samples. The linear model is formally defined as

$$y_{\ i} = b_0 + b_1 x_{1i} + \cdots + b_m x_{mi} + \varepsilon_i \tag{1}$$

Namely

$$y = XB + \varepsilon \tag{2}$$

where $X = \begin{bmatrix} 1\, x_{11}\ x_{12}\ \cdots\ x_{1m} \\ 1\, x_{21}\ x_{22}\ \cdots\ x_{2m} \\ \vdots \\ 1\, x_{n1}\ x_{n2}\ \cdots\ x_{nm} \end{bmatrix}$ and $\begin{cases} y = (y_1, y_2, \cdots, y_n)^T \\ B = (b_0, b_1, \cdots, b_m)^T \\ \varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T \end{cases}$

Commit lest square estimation to formula (2) and we can gain the formula as follows:

$$\hat{B} = (X^T X)^{-1} X^T y \tag{3}$$

We achieve the estimation of y by the formula $y = X\hat{B}$ through putting the formula (3) into formula (2) in order to construct the liner membership function with multi-variables that is defined as

$$A(x) = \alpha + c(\hat{b}_0 + \sum_{i=1}^{m} \hat{b}_i\, x_i) \tag{4}$$

Where α and c are gained by experience on the premise of making $A(x) \in [0,1]$. Give the thresholds λ_i, for example, we should gain the value of λ_1 and λ_2 if we would plan to get three classifications, and the first classification is $\{y \mid A(x) \ge \lambda_1\}$, the second one is $\{y \mid \lambda_1 > A(x) \ge \lambda_2\}$, so the third one is $\{y \mid A(x) < \lambda_2\}$. We realize the classification of independent variables in terms of the thresholds λ_i.

3.3 Proposed Model

To the traditional classification based on the fuzzy statistical analysis, we need to get the correlated factors that affected the change of independent variable. Meanwhile, when the number of factors is too many, or correlation is existed between each factor, a series of data process such as projection, rotating and principal component analysis will be done before least square estimation in order to make the factors independent and the number of factors is suitable. It is too complicated and the processed factors can't reflect all of the original information. And the parameters in the traditional methods are selected by experience in some cases as well, so the classification achieved by this method is uncertain, and furthermore, we can't make vertical analysis to estimate the classification. In the modified method, we get the fuzzy region of each classification through expert knowledge. Then, get the approximate distribution through histogram of each fuzzy region in order to gain the fuzzy membership function $\widetilde{A_n}(x)$ based on hypothetical test not through finding the correlated factors. Finally, compute the fuzziest points of each fuzzy membership function through the equation $\widetilde{A_n}(x)=1/2$ and accordingly achieve the classification by the fuzziest points in order to avoid parameters selected by experience. For example, if we want to estimate the level of some cities' industrial growth. Firstly, translate the concept of industrial growth level into a ten level rating system and make scores for each city in order to get the fuzzy regions of industrial growth level in cities. Secondly, gain the data distribution through each fuzzy region to achieve the fuzzy membership function. Finally, compute the fuzziest points to grade the industrial growth level in cities. The steps of modified method put forward in this paper can be summarized in the follows:

Step 1: Gain the number of classification;

Step 2: Gain data regions and concepts of each fuzzy set;

Step 3: Gain fuzzy sample set in terms of expert system;

Step 4: Compute distribution function of fuzzy sample set by statistical analysis, namely membership function $\widetilde{A_n}(x) \quad (n=1,2,\cdots,n)$;

Step 5: Work out all of the fuzziest points x^*_{n1} and x^*_{n2} ($x^*_{n1} \le x^*_{n2}$, $n=1,2\cdots,n$) in each fuzzy sample set by membership functions $\widetilde{A_n}(x)=\frac{1}{2} \quad (n=1,2,\cdots,n)$;

Step 6: Classify mapping data according to all of the fuzziest points, such as

$$\begin{cases} \left[x^*_{n1},x^*_{n2}\right] \ (n=1) \\ \left[x^*_{1},\left(x^*_{12}+x^*_{21}\right)/2\right) \ ,\left[\left(x^*_{12}+x^*_{21}\right)/2,\left(x^*_{22}+x^*_{31}\right)/2\right),\cdots,\left[\left(x^*_{(n-1)2}+x^*_{n1}\right)/2,x^*_{n2}\right] \ (n\ge2) \end{cases}.$$

On the one hand, the proposed method makes full use of prior knowledge of region concept according to each fuzzy sample set gained by expert system in order to make the classification more generic and more objective. On the other hand, the fuzzy membership of each sample is defined by regulation of mapping data distribution, so the classification boundary is gained by data and make the classification more reasonable. Meanwhile, it isn't be affected by the characteristics of mapping data distribution either, so the modified method in this paper can be used widely.

4 Experiments

In the experiment below, we describe the proposed model by make a thematic map of the presentation of “children”, “junior”, “youth”,” middle age”, “old age” [5,6].

Namely $n=5$, $X=[0,100]$, Fuzzy set means “children”, means “junior”, means “youth”, means “middle age” and means “old age”. Each population research expert has their opinion of the age of “youth”, we did a investigation of 129 experts and the results showed in Table 1.

Table 1. The age of youth

18~25	17~36	17~28	18~25	16~35	14~25	18~30	18~35	18~35	16~25
15~30	18~35	17~30	18~25	10~25	18~35	20~30	18~30	16~30	20~35
18~30	18~30	15~25	18~30	15~25	16~28	16~30	18~30	16~30	18~35
18~25	18~25	16~28	18~30	16~30	16~28	18~35	18~35	17~27	16~28
15~28	16~30	19~28	15~30	15~26	17~25	15~36	18~30	17~30	18~35
15~28	18~30	15~25	15~25	18~30	16~24	15~25	16~32	15~27	18~35
16~35	15~25	15~25	18~28	16~30	15~28	18~35	18~30	17~28	18~35
16~25	18~28	16~28	18~30	18~35	18~30	18~30	17~30	18~30	18~35
16~30	18~35	17~25	15~30	18~25	17~30	14~25	18~26	18~28	18~35
18~28	18~30	18~25	16~35	17~29	18~25	17~30	16~28	18~30	16~28
15~30	15~35	18~30	20~30	20~30	16~25	17~30	15~30	18~30	16~30
18~28	18~35	16~30	15~30	18~35	18~35	18~30	17~30	18~35	17~30
15~25	18~35	15~30	15~25	15~30	18~30	17~25	18~29	18~28	

Group by X, we can get $\widetilde{A}_1$ membership frequency about median, the histogram is as Fig.1.

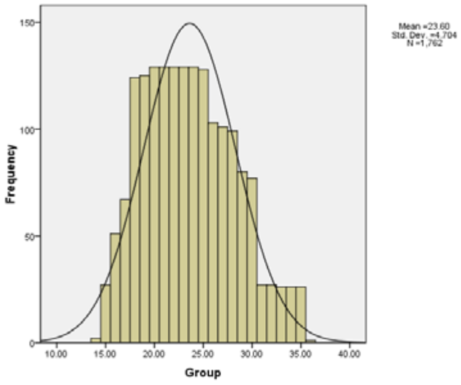
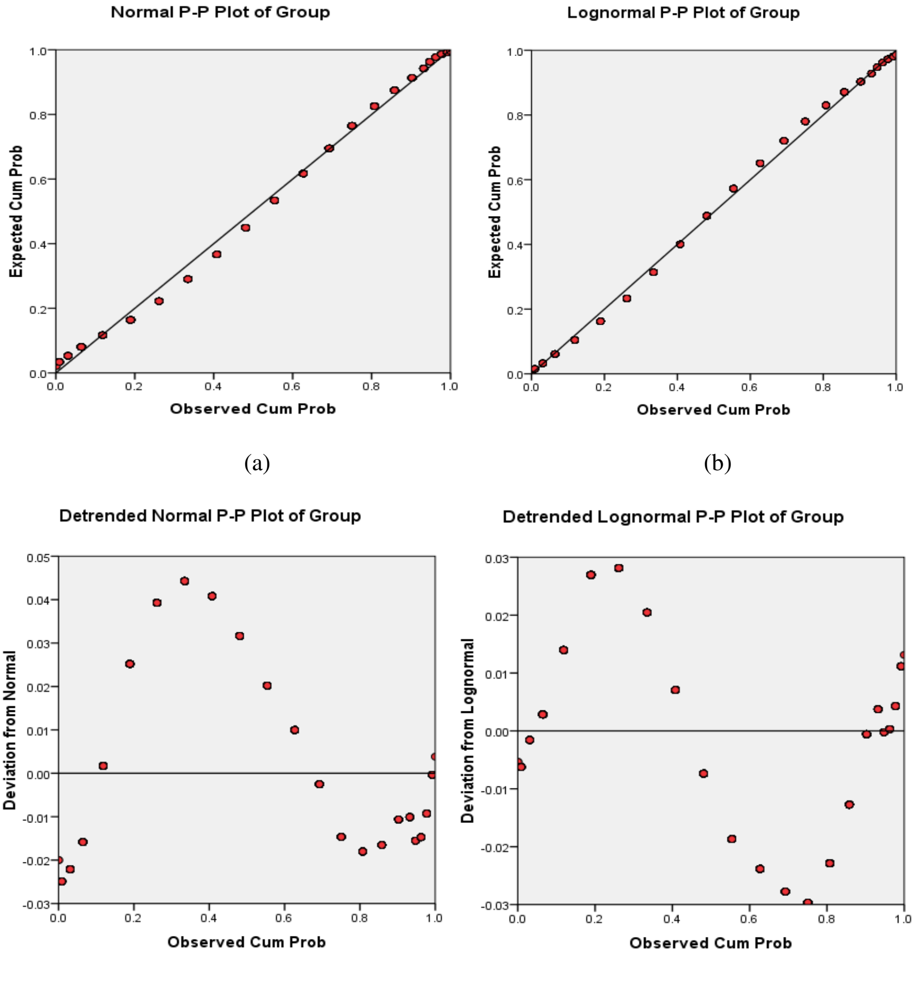


Fig. 1. The histogram of $\widetilde{A}_1$ membership frequency about median

P-P probability map is used to test the distribution [7]. If the data set be tested meet the specified distribution, sample data point will distributed in a straight line. The test of normal distribution and log - normal distribution to grouped data are as Fig.2 (a), (b), (c) and (d).

From Fig. 2 we can know that grouped data meet normal distribution and $\mu_3 = 3.1413, \sigma_3 = 0.19993$, membership function is:

$$\widetilde{A_3(x)} = \frac{1}{\sqrt{2\pi}\sigma_3} e^{-\frac{\left(\ln x - \mu_3\right)^2}{2\sigma_3^2}} \tag{5}$$



(a)

(b)

(c)

(d)

Fig. 2. The test of normal distribution and log - normal distribution to grouped data

Namely $\widetilde{A_3(x)} = \frac{1}{2}$, we can get $x_{31} = 17, x_{32} = 30$, which means the domain of “junior” is $[17,30]$. By the same rule, we can acquire sample data of $\widetilde{A_1}$, $\widetilde{A_2}$, $\widetilde{A_4}$ and $\widetilde{A_5}$ based on the experiments of experts, find the membership function by sample distributes and get domains of them: $[0,10]$, $[8,16]$, $[33,52]$ and $[50,100]$. So the Fuzzy separate domains are $[0,9)$, $[9,17)$, $[17,32)$, $[32,51)$, $[51,100)$. The thematic map of age group rate in each region is designed based on the classification as Fig.3.

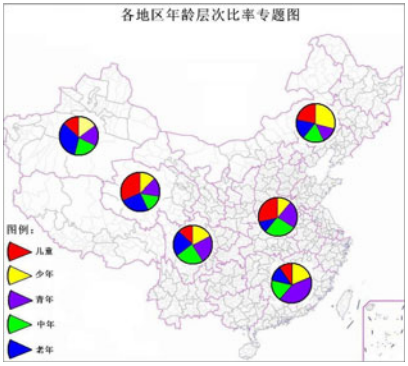


Fig. 3. The thematic map of age group rate in each region

5 Conclusion

The method of mapping data classification based on modified fuzzy statistical analysis is an improvement of traditional multivariate statistical analysis. The proposed method solves the problem of misconstruction of the original data, not being generally used, complexity of computing and fuzzy classification in the traditional mapping data classification method and practical process, and results of classification are objective and reasonable also.

References

1. Hu, B.-q.: Fuzzy Theory Foundation, pp. 271–300. WuHan University Press, Wuhan (2004)
2. Chen, Y.-f., Jiang, N.: Map design Principle, pp. 112–116. Publishing House of PLA (2001)
3. Wang, J.-y., Zhou, J.-h.: The Method to Mapping Data Processing, pp. 168–171. Publishing House of PLA (1992)
4. Lv, A.-m., Liu, X.-t., Guo, J.-z.: An investigation of Classification Rule Based on Bayes Theorem. Application Research of Computers 23(2), 24–25, 72 (2006)
5. Luo, Y.-d., Luo, M.: Application of Rough Set Theory in Spatial Data Mining. Computer and Modernization 2(2), 77–80 (2006)
6. Lv, A.-m., Li, C.-m.: GIS Attribute Data Mining Based on Statistical Inductive Learning. Journal of Institute of Surveying and Mapping 2(3), 290–293 (2001)
7. Lu, W.-d.: SPSS for Windows, pp. 140–160. Publishing House of Electronic Industry (2003)

Web Clustering Using a Two-Layer Approach

Yanping Li, Jinsheng Xing, Rui Wu, and Fulan Zheng

School of Mathematics and Computer
Shanxi Normal University
Linfen, Shanxi 041004, China
{happyyanpingli,jll-li}@163.com
{wurui19710905,zhengfulan668}@126.com

Abstract. Internet is a rich and potential information base. It needs scientific and effective methods in order to find interesting information. Researchers have proposed many web clustering algorithms, but it spends too much time using a simple kind of clustering algorithms, because the number of the web information is huge. Considering the efficiency and the effect of the clustering, in the paper, we use a two-layer web clustering approach to cluster for a number of web access patterns from web logs. At the first layer, we use the LVQ (Learning Vector Quantization) neural network to group the web access patterns to several representative clustering centers. At the second layer, the rough k-means algorithm is adopted to deal with the result of the first layer, producing the final classifications. The experimental results show that the effect is close to monolayer clustering algorithm the rough k-means, and the efficiency is better than the rough k-means by using the two-layer web clustering approach.

Keywords: web clustering, LVQ neural network, rough k-means algorithm, the two-layer approach.

1 Introduction

With the rapid development of the network, there are a growing number of Internet users. We need to analysis and cluster users' access patterns from web logs in order to realize users' visiting activities more conveniently and exactly.

In 1974, Tou and Gonzalez [1] proposed the k-means clustering algorithm, whose classifications are presented by their clustering centers. It has been widely used because of its better effect. Kohonen [2] brought forward the LVQ neural network, which is also important for clustering.

Traditional clustering algorithms make a pattern completely belong to a classification or not. The results of the clustering are crisp classifications, which is unreasonable. So the theories of the fuzzy and rough are used to cluster by more and more researchers. Ruspini [3] gave conceptions of the fuzzy classification for the first time and systematically described the fuzzy clustering. Bezdek [4] proposed fuzzy c-means (FCM) algorithm, which uses the theories of the fuzzy set to group n patterns to c clustering centers with different subjections.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 211–218, 2011.

Lingras P [5] brought the theories of the rough to data clustering and proposed rough clustering algorithm. De [6] applied rough approximation method to cluster users' access patterns from web logs.

In order to improve the efficiency of the clustering, we use a two-layer web clustering approach. LVQ neural network is used to group users' access patterns from web logs to several representative clustering centers. In addition, the rough k-means algorithm clusters for the result of LVQ. Using the approach, users' access patterns can be clustered more effectively.

2 Related Theories of the Fuzzy and Rough

Definition 1. The membership function [7] μF is defined as a function from domain U to interval $[0,1]$, where F is a fuzzy subset. The membership function μF shows the degree how much μ belongs to F .

Definition 2. Linguistic Variable [7] is defined as a multi-component $(x, T(x), U, G, M)$, where x is a variable name, $T(x)$ is the word set of x , U is the domain and G is the grammar rule. In the paper, put users' access time t as a linguistic variable and its word set $T(t)$ is represented by

$$T(t) = \{short, middle, long\}.$$

Definition 3. According to Pawlak [8], upper approximation and lower approximation of $X \subseteq U$ defined as follows.

$$\underline{R}X = \{x \in U \mid [x] \subseteq X\}$$
$$\overline{R}X = \{x \in U \mid [x] \cap X \neq \phi\}$$

Here $[x]$ is the equivalence classification which includes the element x in the relation R .

Definition 4. According to Pawlak [8] X is rough in relation R if and only if $\overline{R}X \neq \underline{R}X$, otherwise we say X is identifiable. So the rough set of X is defined as an interval $A_R(X) = (\underline{R}X, \overline{R}X)$ .All the elements in $\underline{R}X$ belong to X . But the elements in $\overline{R}X$ may belong to X or may not.

3 Data Pretreatment

If a user visits a web page, it means that he is interested in the web page. And the longer time a user stays on a web page, the higher degree interest the user has. Thus, we use the two-dimensional array $(url \;, t \;)$ to present a user's access pattern. There, url is the name of the web page which is visited by a user. And t is the time stayed on the related web page url .

Definition 5. Url is a set of web pages visited by users.
$Url = \{url_1, url_2, ..., url_m\}$, where m is the number of web pages.

Definition 6. S is a set of users' access patterns
$S = \{s_1, s_2, ..., s_n\}$, where n is the number of users' access patterns.

Definition 7. s_i is the ith user's access pattern.

$$s_i = \{(url_{i1}, t_{i1}), (url_{i2}, t_{i2}), ..., (url_{ip}, t_{ip})\}$$

$1 \le p \le m$, $url_{ij} \in Url$, $1 \le j \le p$, where t_{ij} is the ith user's time stayed on the web page url_{ij}.

3.1 Equivalent Dimension Treatment

Users' access activities happen according to their interests. So the length of users' access patterns is different to each other. In order to calculate more conveniently, we represent a users' access pattern as an equivalent dimension vector using (1).

$$h_{ij} = \begin{cases} t_{ij} & if((url_{ij}, t_{ij}) \in s_i) \\ 0 & otherwise \end{cases} \tag{1}$$

The ith user's access pattern is represented as follows.

$$s_i =< h_{i1}, h_{i2}, ..., h_{im} >$$

3.2 Fuzzy Treatment

We cannot exactly describe users' access patterns and experts have certain subjectivity when they make decisions. So we choose a triangle membership function (Fig. 1) to establish the corresponding relationship between numerical value and linguistic variable. Finally a users' access pattern s_i is represented as f_i.

$$\begin{aligned} f_i =< & \mu_{short}(h_{i1}), \mu_{middle}(h_{i1}), \mu_{long}(h_{i1}), \\ & \mu_{short}(h_{i2}), \mu_{middle}(h_{i2}), \mu_{long}(h_{i2}), ..., \\ & \mu_{short}(h_{im}), \mu_{middle}(h_{im}), \mu_{long}(h_{im}) > \end{aligned}$$

Here $x \in T(t)$, $1 \le j \le m$ and $\mu_x(h_{ij})$ is the triangle membership function which is showed in Figure 1.

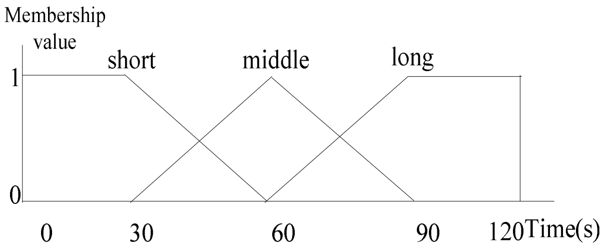


Fig. 1. Membership function of time durations

And the distance between two equivalent dimension vectors is defined as (2).

$$d(f_i, f_j) = \frac{\sqrt{\sum_{k=1}^{m} {}_{x\in T(t)} (u_x(h_{ik}) - u_x(h_{jk}))^2}}{m} \tag{2}$$

4 Web Access Patterns Clustering Based on LVQ Neural Network and Rough k-Means

In the two-layer web clustering approach, LVQ neural network is adopted, which can reduce the time complexity of the whole clustering and increase the efficiency. But the LVQ neural network is sensitive to initial parameters, which makes the result of the clustering is unsatisfactory. So we combine the LVQ neural network with the rough k-means to improve the effect of web clustering.

4.1 Web Clustering Using LVQ Neural Network

LVQ (Learning Vector Quantization) neural network is proposed by Kohonen, which can be used to cluster for web access patterns. LVQ is a winner-take-all and unsupervised neural network, which is composed of the input layer and the output layer. Its network structure shows as Figure 2.

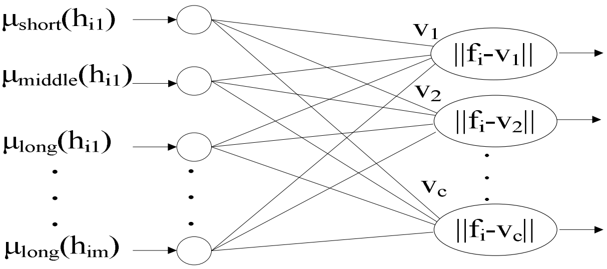


Fig. 2. LVQ neural network structure

The swatch is the set of users' access patterns S and n is the number of the swatch. The number of the input layer is $3m$ .The number of the output layer is c that is also the number of clustering centers. The connection between input layer neuron and output layer neuron is absolute. The set $V = \{v_1, v_2, ..., v_c\}$ is used to describe weights between them. After training, the victorious neural will output 1, others will output 0.

Algorithm 4.1. web clustering using LVQ neural network
Input the set of users' access patterns S
Output c clustering centers $v_i, 1 \le i \le c$
Step 1. Choose c users' access patterns for the initial clustering centers $v_{l,0}, 1 \le l \le c$ .Set the number of clustering centers c ,the error value ε ($\varepsilon > 0$)and learning factor $\alpha_0 \in (0,1)$ .Select a value for the maximum number of iterations $\eta_{\max}$.
Step 2. For η=1 to $\eta_{\max}$

For $k = 1$ to n

- Find the victorious neuron to f_k according to (3).

$$\| f_k - v_{i,\eta-1} \|^2 = \min_{1 \le j \le c} \{ \| f_k - v_{j,\eta-1} \|^2 \} \ . \tag{3}$$

- Update the weight of the victorious neuron using (4).

$$v_{i,\eta+1} = v_{i,\eta} + \alpha_\eta h_{i,j,k} (f_k - v_{i,\eta}) \ . \tag{4}$$

Here $h_{i,j,k} = \begin{cases} 1 & if\,(i \ne j) \\ 0 & if\,(i = j) \end{cases}$.

Before updating the synaptic weight, we need to save the value of $v_{i,\eta-1}$ for the following calculating.

- Calculate the error value by (5).

$$E_\eta = \| V_\eta - V_{\eta-1} \| = \sum_{k=1}^{n} \sum_{j=1}^{c} | v_{j,k,\eta} - v_{j,k,\eta-1} | \, . \tag{5}$$

If $E_\eta \le \varepsilon$ stop Else

$$\alpha_\eta = \alpha_{\eta-1} (1 - \frac{\eta}{\eta_{\max}}) \ \cdot \tag{6}$$

Next η

The time complexity of LVQ neural network is $O(nk)$.

4.2 Web Clustering Using Rough k-means Algorithm

If a users' access pattern exactly belongs to a classification, we will put it to the lower approximation of the classification. If not, we will put it to its upper approximation. It is a soft clustering and the results will be presented by rough sets, which is more reasonable and understandable.

Algorithm 4.2. web clustering by rough k-means algorithm

```
Input c clusters $v_i$ ($1 \le i \le c$) produced by LVQ neural network above
Output K clustering results
Step 1. Set the number of the clustering $K$ and the error value $\varepsilon$ ($\varepsilon > 0$), the value
of $w_l$ and $w_u$ .Select $K$ ($1 \le K \le c$) initial clustering centers $V_i$ ($1 \le i \le K$).
Step 2. For $i = 1$ to $c$
    Set the value of the minimum distance $\min$ .
    For $j = 1$ to $K$
        • Calculate the distance between the $ith$ user's access pattern and the $jth$
          clustering center, which is expressed by $d(v_i, V_j)$.
        • Compare $d(v_i, V_j)$ with the minimum distance $\min$ .
    If $d(v_i, V_j) \le \min$ then $\min = d(v_i, V_j)$.
    Next $j$
    For $j = 1$ to $K$
    For $q = (j+1)$ to $K$
    If $(d(v_i, V_j) - d(v_i, V_q)) < \varepsilon$
    Then $v_i \in \overline{RE}_j$ ▯ $v_i \in \overline{RE}_q$ and $v_i$ don't belong to the lower approximation of
any fuzzy set.
    Else If $d(v_i, V_j) == \min$ then $v_i \in \underline{RE}_j$ .
    If $d(v_i, V_q) == \min$ then $v_i \in \underline{RE}_q$ .
    Next $q$ Next $j$ Next $i$
Step 3. Recalculate clustering centers according to (7) and (8).
    If $| \overline{RE}_i - \underline{RE}_i | \ne \phi$
```

$$V_i = w_l \frac{\sum_{v_k \in \underline{RE}_i} v_k}{| \underline{RE}_i |} + w_u \frac{\sum_{v_k \in (\overline{RE}_i - \underline{RE}_i)} v_k}{| \overline{RE}_i - \underline{RE}_i |} \ . \tag{7}$$

Else

$$V_i = w_l \frac{\sum_{v_k \in \underline{RE}_i} v_k}{| \underline{RE}_i |} \ . \tag{8}$$

Here v_i is an equivalent dimension vector, whose length is n. Parameter w_l and w_u respectively indicate the importance of upper approximation and lower approximation of a rough set. The value $|\underline{R}E_i|$ is the number of patterns that are included by the lower approximation of the classification E_i and $|\overline{R}E_i - \underline{R}E_i|$ is the number of patterns which is between the upper approximation and lower approximation of the classification E_i.

Step 4. Repeat step 2 and step 3 until convergence. For example there are not new clustering centers.

The time complexity of rough k-means is $O(nk^2)$.

5 An Experiment

In the experiment, twenty groups of users' access patterns are chose to cluster in order to test and verify the effect and efficiency of the two-layer web clustering approach we propose.

Firstly, data pretreatment is used to handle with the primitive users' access patterns include the equivalent dimension treatment and the fuzzy treatment. Then the users' access patterns become the equivalent dimension fuzzy vectors.

We respectively use rough k-means algorithm and the two-layer web clustering approach to cluster for the handled uses' access patterns.

The efficiency of algorithms is indicated by the time complexity of algorithms $T(n)$ .In the two-layer approach, we use LVQ to cluster for all of users' access patterns, but we adopt rough k-means to cluster for several representative clustering centers producing by the LVQ neural network. The results show as the following table.

Table 1. Time complexity of algorithms

clustering algorithm	$T(n)$
LVQ neural network	$O(nk)$
rough k-means	$O(nk^2)$
our approach	$O(nk)$

And we use Davies-Bouldin clustering validity index [9] to describe the effect of algorithms, which is defined as (9).

$$d = \frac{1}{n}\sum_{j=1}^{n}\max_{i\neq j}\left\{\frac{S(C_i)+S(C_j)}{d(C_i,C_j)}\right\} . \tag{9}$$

Here $S(C_i)$ is the within-cluster separation and $d(C_i, C_j)$ is the between-cluster distance. The results show as following.

Table 2. Clustering efficiency of algorithms

clustering algorithm	d
rough k-means	0.631
our algorithm	0.597

6 Conclusion

The users' interests are related to the visited page and the time durations on it. Then we put the initial users' access patterns into equivalent dimension fuzzy vectors. The two-layer web clustering approach is composed of LVQ neural network and rough k-means. After training, we see that the effect of the algorithm is close to the monolayer rough k-means clustering algorithm, but the efficiency of the algorithm is better. It can be used to cluster for web access patterns more quickly. And we can provide personalized service for users.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (60574060) and the National Natural Science Foundation of Shanxi province (2006011039).

References

1. Tou, J., Gonzalez, R.: Pattern Recognition Principles. Addison-Wesley, London (1983)
2. Kohone, T.: Self-Organization and Associative Memory, 3rd edn. Springer, Heidelberg (1989)
3. Ruspini, E.H.: A new approach to clustering. Information and Control 19(15), 22–32 (1969)
4. Bezdek, J.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
5. Lingras, P., West, J.: Interval:Set clustering of Web users with rough k-means. Journal of Intelligent Information
6. De, S., Krishna, P.: Clustering web transactions using rough approximation. Fuzzy Sets and Systems 148 (2004)
7. Pawlak, Z.: Rough sets:Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishers, Dordrecht (1991)
8. Pawlak, Z.: Rough sets. International Journal of Information and Computer Science 11, 341–356 (1982)
9. Bezdek, J., Pal, N.: Some new indexes for cluster validity. IEEE Transactions on Systems, Man, and Cybernetic, Part-B 28, 301–315 (1998)

An Improved KNN Algorithm for Vertical Search Engines

Yubo Jia, Hongdan Fan, Guanghu Xia, and Xing Dong

Institute of Information and electron
Zhejiang Sci-Tech University
Hangzhou, 310018, China
Jiayubo1964@163.com

Abstract. Secondary Data Processing deals the information further by re-crawling and categories based on the basic of structured data. It is the key researching module of Vertical Search Engines. This paper proposes an improved KNN algorithm for the categories. This algorithm achieves the responsiveness and the accuracy of vertical search by reducing the time complexity and accelerating the speed of classification. The experiment proved the improved algorithm has the better feasibility and robustness when it's used in secondary data processing and participle of vertical search engines.

Keywords: KNN, categories, similarity, search, vertical search.

1 Introduction

With the development of the Internet, search engines continue to meet the demand of great information resources, but cannot take into account the accuracy and responsiveness[1] of information search, so vertical search engines have emerged to meet the needs of users at this time. But how to classify web pages and texts in the searching process is critical. In this paper, we propose an improved KNN algorithm to reach the aim of categories.

2 KNN Categories Algorithm

A. The Basic Idea of KNN Categories Algorithm

To represent the texts that need to be classified and web pages as vectors and calculate the similarity of vectors between web pages and samples from the space consisted by the training samples. Then we can obtain k pieces of nearest and most similar texts or pages. According to the type of these web pages to determine the category of the new ones, then compute the seniorities of classes in the k neighbors from the new version in turn. Assign the pages and texts to the class with the most powerful seniority.

Texts are represented into a vector in the vector space model, so calculating the similarity of texts can be transformed into computing the cosine of the angle between vectors. Assume two pages or texts d_i = (w_{i1} , w_{i2} ,…, w_{in}), d_j = ($w_{j1}, w_{j2}, \dots, w_{jn}$), The formula of similarity between d_i and d_j is $sim(d_i, d_j)$.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 219–223, 2011.

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{i=1}^{n} w_{ik}^2}\sqrt{\sum_{j=1}^{n} w_{jk}^2}}$$

The greater value of $sim(d_i, d_j)$ means the smaller angle between d_i and d_j. When $sim(d_i, d_j) = 1$, d_i and d_j are parallel or coinciding. At this time they are most similar. If $sim(d_i, d_j)$ is near 0, it means vectors d_i and d_j are vertical. They have the smallest value of similarity.

We can compute the seniorities of classes in the k neighbors from the new version in turn. The formula of seniorities is $P(\bar{x}, C_j) = \sum_{d_{i \in KNN}} sim(\bar{x}, \bar{d}_i) y(\bar{d}_i, C_j)$.

$\bar{x}$ is the basic characteristic quantity of new pages. $sim(\bar{x}, \bar{d}_i)$ is the formula of similarity. $y(\bar{d}_i, C_j)$ is type property function, it means if class C_j includes $\bar{d}_i$, then $y(\bar{d}_i, C_j)$=1, else $y(\bar{d}_i, C_j)$=0.

B. Analysis of Common KNN Algorithm[2]

The advantage of the traditional KNN arithmetic is that it can make use of the correlation directly between the two given samples, thus reducing the influence caused by inadequate choices from classification feature and also the error term during the process of classification. But this algorithm compares with every sample vector in sample space in order to find k neighbors of sample classification, so it causes computing times to increase and system performance degrades.

3 Improved KNN Algorithm

A. The Basic Idea of Improved KNN Categories Algorithm

When searching k neighbors of one sample, only find those which have overlapping words with those unsorted pages, thus reducing the search scope and accelerating the speed of search.

The structure of the improved algorithm includes term arrays and their lists. Term array is the ID of feature entry that stored in arrays undergone the feature extraction after dividing the training texts into words.

Every entry（ti）in term arrays has its own pointer, which points to the list formed by all of the texts. The text list includes two parts, such as ID and the seniorities ti in text. After list of the text including ti is finished, sorting decreasingly according to the value of seniorities, then make a further optimization to narrow the search range of KNN algorithm.

B. Description of Improved KNN Categories Algorithm

Page d ready to be sorted is expressed as text vector V(w_1, w_2,… ,w_n), search each document list li (1≤i≤n) of term ti(1≤i≤n) in vector V, then merge list li and remove the ID of same texts in lists, so we can obtain the similarity between the set of texts ID and texts in the set.

3.1 Analysis of Improved KNN Algorithm

Similarity is only between the improved algorithm and the documents vectors of intersectional training texts ready to category. So it can reduce the time complexity[3] and accelerate the speed of classification on a certain extend. But the improved KNN algorithm is more similar with part of training samples, so there are many certain overlaps in sample vectors. The improved algorithm is a compromised algorithm compared with common algorithm and has a better practicability.

4 Application of Improved Algorithm in Vertical Search Engines

A. Theoretical Description

Double Data Processing and Participle Module are the key modules of the model. It is a process of structured analysis that the mode separation of the text content on web pages stored in the database, adjustment of data[4] and analysis of related links[5]. It will deal the information further by re-crawling and categories based on the basic of structured data. This improved KNN classification algorithm stresses the responsiveness and the accuracy of vertical search when it's used in vertical search engines model.

B. Experiments

This experiment designs two evaluation indicators, precision ratio and recall ratio, for the improved KNN algorithm. Precision ratio[6] is defined as the percents of texts which meet with the result of artificial classification in all given documents. Its mathematical formula is Precision ratio=(the number of correct classification texts/all classification texts). Recall ratio is defined as the rate of texts which meet with classification system in all the deserved results of artificial classification. Its mathematical formula is Recall rate=(a mount of correct classification texts/ deserved texts). Precision rate and recall rate reflect the quality of classification in two different aspects. We should balance the two aspects. Thus there is a new evaluation indicator, the value of F1, and its mathematical formula is F1[7]=(2*precision rate*recall rate)/deserved texts.

This experiment conducts a test of 300 web pages of the given vertical search engine model. In the KNN algorithm experiment, we set the threshold [8] k=100 and the vectors dimension threshold V are given different values.

The experiment results are showed as figure 1.

C. Experimental results and analysis

The precision and recall of KNN Algorithm are showed as figure 1 and figure 2.

It can be seen from figure 1 obviously, the improved KNN algorithm is slightly more common in precision, but when the dimensions of vectors reached a certain value, the accuracy decreased significantly. So in the case of certain threshold, the vector dimension should be appropriate. We can discover from figure 2 that recall is less affected by the dimension of vectors and keeps balance overall, but recall of our improved KNN algorithm increases clearly.

Table 1. Experimental Results

	Precision （%）		Recall （%）	
	Common KNN	Improved KNN	Common KNN	Improved KNN
V=10	80.21	80.78	80.45	92.90
V=15	81.38	82.13	83.47	95.32
V=20	85.63	86.14	85.92	96.13
V=25	82.12	83.07	85.84	95.46
V=30	81.08	80.56	84.35	95.38

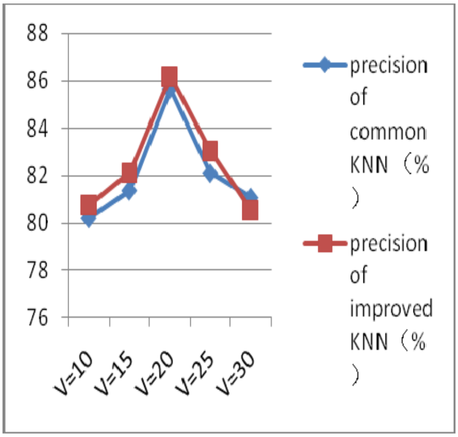


Fig. 1. Precision of KNN Algorithm

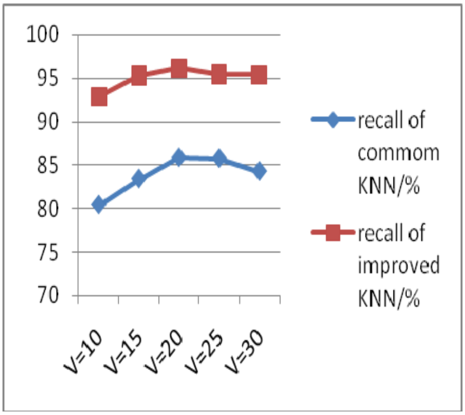


Fig. 2. Recall of KNN Algorithm

5 Conclusions

This improved KNN algorithm enhances the recall and efficiency of categories significantly on the basis of slight loss of precision. In the vertical search engines, improved KNN algorithm can achieve classification function of the secondary processing and Participle n module. Therefore, the improved algorithm has a better practicability for vertical search engines.

References

[1] Yang, J., Ling, P.: Improvement of PageRank Algorithm for Search Engine. Computer Engineer 35(22), 35–37 (2009)

[2] Pan, L., Yang, B.: Study on KNN Arithmetic Based on Cluster. Computer Engineering and Design 30(18), 4260–4261 (2009)

[3] Yan, W., Wu, W.: Data Structures(C Edition), pp. 13–17. Tsinghua University Press, Beijing (2008)

[4] Soderland, S., Cardie, C., Mooney, R.: Learning Information Extraction Rules for Semi-strucrured and Free Text. Machine Learning (1999)

[5] Bertoli, C., Crescenzi, V., Merialdo, P.: Crawling Programs for Wrapper-based Application. In: IEEE IRI 2008, pp. 160–165 (2008)

[6] Zhang, N., Jia, Z., Shi, Z.: Text Categorization with KNN Algorithm. Computer Engineering 31(8), 171–185 (2005)

[7] Pei, Z., Shi, X., Maurizio, M., Liang, Y.: An Enhanced Text Categorization Method Based on Improved Text Frequency Approach and Mutual Information Algorithm. Progress in Natural Science 17(12), 1494–1500 (2007)

[8] Shao, F., Yu, Z.: Principle and Algorithm of Data Mining, pp. 126–176. Waterpub Press, Beijing (2003)

Concluding Pattern of Web Page Based on String Pattern Matching

Yiqing Cai[1], Xinjun Wang[1], Chunsheng Lu[2], Zhongmin Yan[1], and Zhaohui Peng[1]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] Information Center of Ministry of Human Resources and Social Security of the People's Republic of China
caiyiqing1987@163.com, {wxj,yzm,pzh}@sdu.edu.cn, luchunsheng@mohrss.gov.cn

Abstract. Presently, each Web site has its own topics and formats to arrange the page structure and present information. Therefore, there is a great need for value-added service that extracts information from multiple sources. Data extraction from HTML is usually performed by software modules called wrappers. In many studies of constructing wrapper, concluding the pattern of the Web site is a importance task in the beginning. This paper studies the problem of concluding pattern from a Web page that contains several nested structure and repeated structure. In our method, the algorithm bases on string pattern matching can discover the nested structure and the repeated structure in a Web page. Then a regular expression will be generated as the pattern of the Web site.

Keywords: hierarchical preorder traversal string, pattern of Web site, string pattern matching, nested structure, the repeated structure, concluding pattern.

1 Introduction

At present, each Web site has its own topic and formats to arrange the page structure and present information. Therefore, there is a great need for value-added service that extracts information from multiple sources. Data extraction from HTML is usually performed by software modules called wrappers. In many studies of constructing wrapper, concluding the pattern of the Web site is one of the most importance tasks.

In this paper, we propose a novel and effective method to conclude pattern in a Web page automatically. The algorithm bases on string pattern matching can discover the nested structure and the repeated structure in a Web page automatically.

Typically, every record in Html page is corresponds to an entity in real world. These records are represented by multiple data elements together in a particular mode of organization. They are semi-structured data. Fig. 1 is a HTML page from amazon.com. It contains much information about more than one book. In this page, each book is showed in the form of Web data record. Each record all contains some data elements associated with the book, such as book name, author, publishing company, introduction, and so on. It is obvious that these records are showed in the same pattern, the data items are organized with certain regular pattern.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 224–233, 2011.

Fig. 1. An example: www.amazon.com

Usually, HTML-based sites contain large amounts of data and a fairly regular structure. It can be found by observing that the same type of records is showed with similar HTML codes in a certain data region of the page. Adding some records to the page or adding some data items to the record can be considered as adding some given HTML codes to given location of the page. And the corresponding block of similar records usually has the same parent in DOM tree. In Fig. 2, we show an instance.

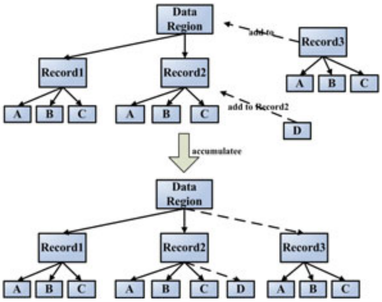


Fig. 2. The structure of web data

Usually, there is some nested type[3] in HTML page. If O1, O2, ..., On are all nested types, then their ordered list <O1, O2, ..., On> is also a nested type.

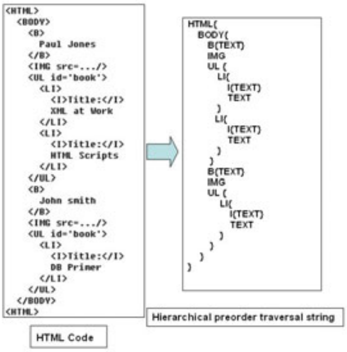
- If O is a nested type, then the set, {O}, is also a nested type.
- If O is a nested type, then (O)? represents the optional type O.
- If O1 and O2 are nested type, then (O1|O2) represents the disjunction of O1 and O2.

Our method has four steps to conclude pattern in a Web page automatically:

(1) Obtain hierarchical preorder traversal string from the DOM tree of web page.
(2) Group the repeated structure by using the hierarchical preorder traversal string.
(3) Reduce the hierarchical preorder traversal string to a regular expression.
(4) Conclude patterns from a certain number of page, then fusion them.

2 Obtain Hierarchical Preorder Traversal String of DOM Tree

We can obtain traditional preorder traversal string of DOM tree easily. But we need a kind of preorder traversal string which is hierarchical. So we transform traditional preorder traversal string as Fig. 3. In hierarchical preorder traversal string, tag TEXT map to text fields. And we use symbol {} to mark hierarchy between all tags. There is only one hierarchical preorder traversal string for a DOM tree. In turn, it is also established.



```
<HTML>
  <BODY>
    <B>
      Paul Jones
    </B>
    <IMG src=.../>
    <UL id='book'>
      <LI>
        <I>Title:</I>
        XML at Work
      </LI>
      <LI>
        <I>Title:</I>
        HTML Scripts
      </LI>
    </UL>
    <B>
      John smith
    </B>
    <IMG src=.../>
    <UL id='book'>
      <LI>
        <I>Title:</I>
        DB Primer
      </LI>
    </UL>
  </BODY>
<HTML>
```

```
HTML{
  BODY{
    B(TEXT)
    IMG
    UL {
      LI{
        I(TEXT)
        TEXT
      }
      LI{
        I(TEXT)
        TEXT
      }
    }
    B(TEXT)
    IMG
    UL {
      LI{
        I(TEXT)
        TEXT
      }
    }
  }
}
```

Fig. 3. Hierarchical preorder traversal string

3 Repeated Structure Clustering

We have mentioned two discovery in the previous section: (1) the same type of records are showed with similar HTML codes in a page; (2)the corresponding block of similar records usually has the same parent in DOM tree. Therefore, we can consider that similar records are organized by adjacent repeated structure usually.

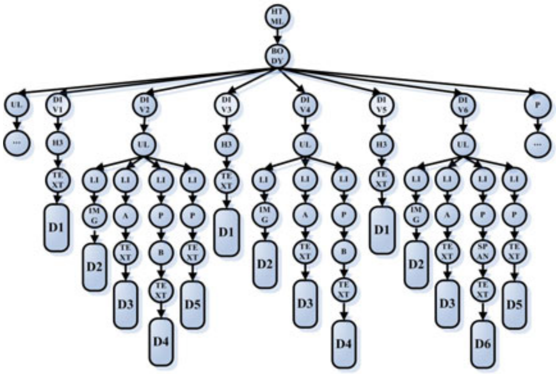


Fig. 4. DOM tree of a web page

In this paper, we define those HTML code blocks, which have Consistent Trees structure or Similar Trees structure, as repeated structure. In this step, we will look for repeated structure by comparing subtree structure of all sibling nodes with the same tag name. Here, we call the subtree whose root is node X as subtree of node X in short. To further explain Consistent Tree structure and Similar Tree structure, we make use of an artificial tag tree in Fig. 4. For explanation convenience, we use ID numbers to denote tag DIV.

3.1 Consistent Trees and Similar Trees

From Fig. 4, we can obtain the hierarchical preorder traversal strings of the subtree of tag DIV2 and tag DIV 4.

DIV2: DIV{UL{LI{IMG}LI{A{TEXT}}LI{P{B{TEXT}}}LI{P{TEXT}}}}①
DIV4: DIV{UL{LI{IMG}LI{A{TEXT}}LI{P{B{TEXT}}}}}②

The difference between them is only the string “LI{P{TEXT}}” of ①. And this string “LI{P{TEXT}}” is also a hierarchical preorder traversal string. We define such sort of two trees as Consistent Trees. String ② is contained in string ①, so we call the subtree structure of tag DIV2 contain the subtree structure of tag DIV4. And we can express this relation by expression: S(DIV4)<S(DIV2)

“LI{P{TEXT}}” is an optional part of a nested type. The subtree structure of tag DIV2 and tag DIV4 can be expressed as following:

DIV{UL{LI{IMG}LI{A{TEXT}}LI{P{B{TEXT}}}(LI{P{TEXT}})?}}

From Fig. 4, we can find the hierarchical preorder traversal string of the subtree of tag DIV1 and the hierarchical preorder traversal string of the subtree of tag DIV3 are same. So the subtree with root tag DIV1 and the subtree with root tag DIV3 is also Consistent Trees. And we can express this relation by expression: S(DIV1)=S(DIV3).

In HTML page, the Consistent Trees probably contains some web records in the same format and type. The different parts probably show an optional data items.

From Fig. 4, we can obtain the hierarchical preorder traversal string of the subtree of tag DIV2 and tag DIV 6.

DIV2: DIV{UL{LI{IMG}LI{A{TEXT}}LI{P{B{TEXT}}}LI{P{TEXT}}}}①
DIV6: DIV{UL{LI{IMG}LI{A{TEXT}}LI{P{SPAN{TEXT}}}LI{P{TEXT}}}}②

The difference between them are the string “P{B{TEXT}}” of ① and the string “P{SPAN{TEXT}}” of ②. The two strings are also hierarchical preorder traversal strings. And the location of “P{B{TEXT}}” in ① is corresponding to the location of “P{SPAN{TEXT}}” in ②. We define such sort of two trees as Similar Trees. And we can express this relation by expression: S(DIV2)≈S(DIV6).

“P{B{TEXT}}” and “P{SPAN{TEXT}}” is a disjunction part of a nested type. The subtree structure of tag DIV2 and tag DIV6 can be expressed as following:

DIV{UL{LI{IMG}LI{A{TEXT}}(LI{P{B{TEXT}}}|
P{SPAN{TEXT}})LI{P{TEXT}}}}

In HTML page, the Similar Tree probably contains some web records in the same format and type. The different parts probably show a disjunctive data items.

In some practical cases, some two tree accord to the standard of Consistent Trees or Similar Trees, but the amount of the different points is too big. We can consider they are not Consistent Trees or Similar Trees. The amount of the different points can be set by practical cases.

3.2 Clustering Algorithm

This arithmetic will visit all nodes of DOM tree by breadth-first search, then deal with the children node of the current node which has the same tag name in two stages. In the first stages, the arithmetic will divided those nodes into different group by Consistent Tree compare. In the second step, the arithmetic will unite those groups, which are reserved in the first step by Similar Tree compare. Through the above two steps, the nodes in the same group have repeated subtree structure. The algorithm for repeated structure clustering is given as follows.

Algorithm. Repeated structure clustering

```
Input: DOM tree of a HTML page
Output: The clustering result of all sibling nodes
 Visit all nodes of DOM tree by breadth-first search
   if current node have children nodes then
     Divide the node having the same tag name into
        the same Set
     for all Seti of Sets
       Executive Function Clustering($Set_i$)
   end if
 Until all nodes have been visited, all repeated structure
 have be found
```

Function. Clustering(Set)

```
// Following is the first stage, Consistent Tree compare
Divide all $node_i$ of Set into $Group_i$ and set $node_i$ as the
maxNode of $Group_i$ $(1 \le i \le n)$
for (i=1; i<n; i++)
  if Groupi hasn't be unite to other Group then
    for (j=i+1; j≤n; j++)
      if Groupj hasn't be unite to other Group then
        if $S(maxNode_i) < S(maxNode_j)$ then
          unite Groupi to Groupj
        else
          if $S(maxNode_i) > S(maxNode_j)$ then
            unite $Group_i$ to $Group_j$
          and set $maxNode_j = maxNode_i$
          end if
        end if
      end if
    end for
  end if
end for
//Following is the second stage, Similar Trees compare
Renumbered all reserved Group after the above step Assign
a $maxNodeSet_i$ to all $Group_i$ and put $maxNode_i$ into the
$maxNodeSet_i (1 \le i \le n)$
for (i=1; i<n; i++)
  if Groupi hasn't be unite to other Group then
```

```
      for (j=i+1; j≤n; j++)
        if Groupj hasn't be unite to other Group then
          if a $maxNode_r$ in $maxNodeSet_i$
         and a $maxNode_s$ in $maxNodeSet_j$
         $S(maxNode_s) \approx S(maxNode_r)$ then
            unite Groupi to $Group_j$
                  and unite $maxNodeSet_i$ to $maxNodeSet_j$
          end if
        end if
      end for
    end if
  end for
```

In the stage of Consistent Trees compare, maxNode denote the node whose has the maximal subtree structure in the Group. In any Group, it have S(node)< S(maxNode) for any node of current Group.

In the stage of Similar Tree compare, maxNodeSet is used to store maxNodes of all the Groups which are united to current Group. It any Group, it haven't S(node)> S(maxNode) for any node and maxNode of current Group. And for any maxNode_r and maxNode_s of maxNodeSet of a Group, it have $S(\text{maxNode}_r) \approx S(\text{maxNode}_s)$.

4 Concluding Pattern

In this step, the hierarchical preorder traversal strings of Html page will be reduced to a regular expression base on repeated structure clustering. Regular expression can be employed to model the nested and repeated structure. Given an alphabet of symbols ∑ and a special token "TEXT" that is not in Σ, a regular expression over ∑ is a string over ∑∪{TEXT, *, ?, |, (,)} defined as follows:

- The empty string ∈ and all elements of ∑∪{TEXT} are regular expressions.
- If A and B are regular expressions, then AB, (A|B) and (A)? are regular expressions, where (A|B) stands for A or B and (A)? stands for (A|∈).
- If A is a regular expression, (A)* is a regular expression, where (A)* stands for ∈ or A or AA or …

SO we can reduce the HTML code of Fix. 5 to regular expression as follows:

HTML{BODY{UL{…} (DIV{H3{TEXT}} DIV{UL{LI{IMG}LI{A{TEXT}}
LI{P{(B{TEXT}|SPAN{TEXT})}}(LI{P{TEXT}})?}})* UL{…}}}

The array of repeated structure has two status:

(1) One kind of repeated structure successive appear, e.g. AAA->A*

(2) Some kinds of repeated structure arrange with certain regulation, then successive appear, e.g. ABCABCABC-> (ABC)*

When we conclude regular expression, we need to estimate the array status of repeated structure. The algorithm for how to estimate is given as follows.

Algorithm. estimate array status

```
Input: Hierarchical preorder traversal strings
Output: Maximal repeated substring
Visit all nodes of DOM tree by breadth-first search
if current node have children nodes then
  for all ChildrenNode$_i$
    if ChildrenNodei be assigned to a Group in
       Repeated structure clustering then
      Mark ChildrenNode$_i$ with tag name and
      the num of Group of repeated structure
      clustering
    else Mark ChildrenNode$_i$ with tag name
    end if
  end for
end if
join all Marking of ChildNodes to a string S in the order
of DOM tree
search as long as possible repeated substring in S
```

5 Maximal Pattern Fusion

If we only conclude pattern from one page, the pattern may have some shortage. Because some structure may not exist in some page. In order to get the right and whole pattern, we need to conclude pattern from a certain number of page, then compare these patterns, and fusion them. The process of fusion patterns is actually a string processing constructing a string contains all patterns.

6 Experiment

Based on the four steps described above, we have developed a prototype of the pattern generation system and used it to run a number of experiments on real HTML sites.

We examined three categories of web sites, book shopping, job advertisements and car advertisements, and collected nine web sites for each category. Table 1 show the

Table 1. Special symbols contained in Web site

	Book			**Job**			**Car**		
	?	**|**	*****	**?**	**|**	*****	**?**	**|**	*****
Site1	Y	N	Y	Y	N	Y	Y	N	Y
Site2	Y	Y	Y	Y	Y	Y	Y	Y	Y
Site3	Y	N	Y	Y	Y	Y	N	N	Y
Site4	Y	Y	Y	N	N	Y	Y	N	Y
Site5	N	N	Y	N	N	Y	Y	Y	Y
Site6	Y	Y	Y	Y	N	Y	Y	N	Y
Site7	N	N	Y	Y	Y	Y	N	N	Y
Site8	Y	Y	Y	Y	N	Y	Y	Y	Y
Site9	N	N	Y	N	N	Y	Y	N	Y

special symbols of "?", "|" and "*" contained in the final pattern (the regular expression) for each web site, from which we can see that web sites do model their data region, navigation and adcolumn both in plain-structure and nested-structure.

In order to test the astringency of fusing pattern, we use sample sets of one site in some different size form each category of web sites to conclude pattern.

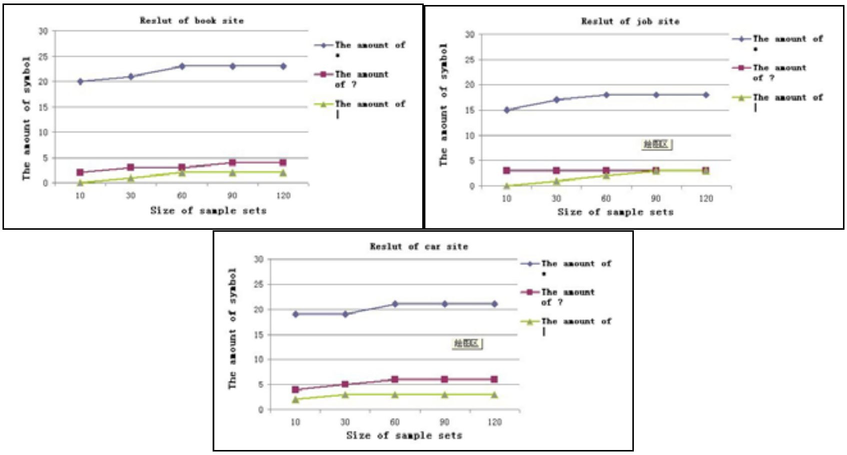


Fig. 5. Astringency of concluding pattern

Table 2. Precision of different methods

	Used in RoadRunner			**Used in Dela**		
	P(%)	**R(%)**	**F(%)**	**P(%)**	**R(%)**	**F(%)**
Site1	82.7	79.4	79.4	92.6	94.3	94.3
Site2	83.6	81.4	81.4	94.8	92.5	92.5
Site3	81.2	78.3	78.3	90.6	91.2	91.2
Site4	85.3	82.6	82.6	93.8	95.7	95.7
Site5	71.2	79.4	79.4	89.6	92.5	92.5
Site6	68.7	62.3	62.3	86.5	84.8	84.8
Site7	83.3	78.5	78.5	91.8	89.1	89.1
Site8	76.6	79.5	79.5	88.5	92.3	92.3
Site9	69.5	71.7	71.7	81.6	83.5	83.5
	RoadRunner			**Dela**		
	P(%)	**R(%)**	**F(%)**	**P(%)**	**R(%)**	**F(%)**
Site1	77.3	69.6	69.6	87.3	90.8	90.8
Site2	85.7	78.6	78.6	92.6	89.4	89.4
Site3	76.5	72.3	72.3	89.8	90.5	90.5
Site4	79.6	72.9	72.9	90.3	92.1	92.1
Site5	62.4	75.6	75.6	86.5	90.2	90.2
Site6	65.2	56.3	56.3	83.7	81.9	81.9
Site7	80.2	74.6	74.6	90.7	88.3	88.3
Site8	64.1	68.3	68.3	85.3	89.5	89.5
Site9	59.8	68.6	68.6	79.6	82.9	82.9

Then we observe stability of the pattern by observing the change of the amount of the special symbols of "?", "|" and "*" contained in the final pattern in different size of sample set of one site. Fig. 5 shows the result of this experiment. We can find that the final pattern will tends to complete as the size of sample set increases.

We reduce the hierarchical preorder traversal strings of Html page to a regular expression to conclude pattern. In this process, actually, we use the special symbols of "?", "|" and "*" to express the repeated and nested structure of the web page. With the implementation of maximal pattern fusion, the amount of the special symbols of "?", "|" and "*" in the final pattern will may increase. When we get the right and whole pattern, the amount of the special symbols will not change. From Table. 2, we can find that the final pattern tends to steady.

Our method of concluding pattern is similar to RoadRunner and Dela. So we put the final pattern of our method in the system of RoadRunner and Dela to generate wrapper. Then we test the precision and the efficiency of our method in the system of RoadRunner and Dela.

In Table. 2, columns labeled by "P" represent the precision of extraction results. Columns labeled by "R" represent the recall of extraction results. Columns labeled by "F" represent comprehensive evaluation of precision and recall------ $F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

From the experimental results, we can find that our method have a good precision in RoadRunner and Dela.

7 Related Work

In order to improve the efficiency and reduce manual efforts, most recent researches focus on automatic approaches instead of manual or semi-automatic ones to analyze the structure of web page and extract data from web page. Some representative automatic approaches are RoadRunner [1], IEPAD [2], DELA[3], MDR [6], DEPTA [4], ViWER[7], VIPER[9], VIDE[11] and the method in [5], [10], [12].there are three approaches based on concluding pattern(regular expression).

Crescenzi et al. develop in [1] a wrapper induction system, ROADRUNNER, which generates a wrapper based on a comparison of the similarities and differences between web pages. This approach can identify nested structures and repeated structures in an HTML page. However, ROADRUNNER assumes the wrappers are union-free regular expression, which cannot catch "the full diversity of structures presented in HTML pages." [1].

Chang et al. propose a system called IEPAD in [2] that generates extraction rules by coding the HTML page into a binary sequence and then mining maximal repeated patterns in the sequence by building a PAT tree (Patricia Tree). This approach is deterministic and efficient for web pages containing plain-structured data objects. However, it cannot handle complex, nested–structured data objects.

Jiying Wang et al. propose a system called Dela in [3]. It assume that data objects contained in HTML pages are generated by some common templates and the structure of embedded data objects may appear repeatedly if the HTML page contains more than one data object instance. The basic idea of their method is based on the iteration of building token suffix-trees and discovering C-repeated patterns.

8 Conclusion

In this paper, we described a novel and effective method to conclude pattern in a Web page automatically. In our method, the algorithm bases on string pattern matching can discover the nested structure and the repeated structure in a Web page. The experimental results also demonstrate the feasibility of discovering the nested structure and repeated structure basing on string pattern matching in a Web page automatically. As future work, we plan to include some visual information into our method which can help us to discover the nested structure and repeated structure. And we plan to do some study of generating wrapper base on our method.

Acknowledgments. Supported by the National Key Technologies R&D Program No.2009BAH44B01, Supported by the National Natural Science Foundation of China under Grant No.61003051, the Natural Science Foundation of Shandong Province of China under Grant No. ZR2010FM033, Key Technology R&D Program of Shandong Province under Grant No. 2010GGX10108.

References

1. Crescenzi, V., Mecca, G., Merialdo, P.: ROADRUNNER: towards automatic data extraction from large web sites. In: Proc. 27th VLDB Conf., pp. 109–118 (2001)
2. Chang, C.H., Lui, S.C.: IEPAD: information extraction based on pattern discovery. In: Proc. 10th Intl. Conf. on World Wide Web, pp. 681–688 (2001)
3. Wang, J., Lochovsky, F.H.: Data extraction and label assignment for web databases. In: Proceedings of the 12th international conference on World Wide Web (2003)
4. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: Proceedings of the 14th International Conference on World Wide Web (2005)
5. Chang, C.-H., Hsu, C.-N., Lui, S.-C.: Automatic information extraction from semi-Structured Web Pages by pattern discovery. Decision Support Systems Journal 35(1), 129–147 (2003)
6. Liu, B., Grossman, R., Zhai, Y.: Mining data records from Web pages. In: KDD (2003)
7. Hong, J.L., Siew, E.-G., Egerton, S.: ViWER-data extraction for search engine results pages using visual cue and DOM Tree. In: International Conference on Information Retrieval & Knowledge Management (CAMP), pp. 167–172 (2010)
8. Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y.: VIPS: a visionbased page segmentation algorithm. Microsoft Technical Report, MSR-TR-2003-79 (2003)
9. Simon, K., Lausen, G.: Viper: augmenting automatic information extraction with visual perceptions. In: CIKM, pp. 381–388. ACM, New York (2005)
10. Li, L.Z., Liu, Y.H., Obregon, A., Weatherston, M.: Visual segmentation-based data record extraction from Web documents. In: IEEE International Conference on Information Reuse and Integration, pp. 502–507 (2007)
11. Liu, W., Meng, X.F., Meng, W.Y.: ViDE: A Vision-based Approach for Deep Web Data Extraction. IEEE Transactions on Knowledge and Data Engineering 22(3), 447–460 (2010)
12. Miao, G., Tatemura, J., Hsiung, W.-P., Sawires, A., Moser, L.E.: Extracting Data Records from the Web Using Tag Path Clustering. In: Proceedings of the 18th International Conference on World Wide Web (2009)

A Rapid Method to Extract Multiword Expressions with Statistic Measures and Linguistic Rules[*]

Lijuan Wang and Rong Liu

Dept. of Computer Science and Software Engineering,
Taiyuan University of Technology,
030024, Taiyuan, China
wanglj@tyut.edu.cn

Abstract. Multiword Expressions (MWEs) have been the bottleneck in NLP. Particularly, the resource of fixed MWEs can improve the performance of tasks and implications of NLP. Due to complex characters of MWEs, it is hard to make difference between fixed MWEs and unfixed MWEs. This paper puts forwards an approach to extract fixed MWEs rapidly. First the definition of fixed MWEs is given. Features contributing to determinate fixed MWEs are considered both in statistic measures and in linguistic information. We extract fixed MWEs in the frame of multi-features and do manual evaluation. Experiment shows that the approach is effective. Our job can provide a desired list of fixed MWEs for NLP implication.

Keywords: Multiword Expressions, Maximum Entropy, Semantic Rules, Stop Words.

1 Introduction

There is no uniform definition of MWEs. The definition of MWEs given by Sag is “any word combination for which the syntactic or semantic properties of the whole expression cannot be obtained from its parts” (Sag et al., 2002). Examples of MWEs are idioms (kick the bucket, rock the boat), phrasal verbs (depend on, go to), compounds (traffic light, police car), etc. MWEs is numerous in languages, which account for between 30% and 45% of spoken English and 21% of academic prose according to Biber et al. (1999), and by Jackendoff (1997) the number of MWEs in a speaker’s lexicon is of the same order of magnitude as the number of single words.

In WordNet, almost half of entries are composed of MWEs. However, these figures are likely to be underestimated if we consider that for language from a specific domain the specialized vocabulary is going to consist largely of MWEs (three represents theory, dual processor) and new MWEs are constantly appearing (climate change, cloud computing). Hence, Helena de Medeiros Caseli et al.(2009) indicates that MWEs in a special domain are likely to be underestimated, if we consider MWEs consist of a number of terms and more new MWEs flow increasingly. MWEs have been the

[*] This research has been funded by Taiyuan University of Technology（Item number: 900103-03010255 and 900103-03020632）.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 234–241, 2011.

bottleneck for high-quality NLP applications, which should not only identify MWEs but also to deal with MWES when they are found. For parsing, Baldwin et al. (2004) found that mistakes due to no identification of MWEs amounts to 8% in the experiment of randomly selecting 20000 sentences in British National Corpus. Therefore a robust method to extract MWEs automatically or semi-automatically is in an urgent need to build language resource for tasks of Natural Language Processing. MWEs play a vital role in Natural Language Processing applications, especially for Machine Translation. Moreover, many fields including word disambiguation, automatic text classification, lexical compilation, information retrieval need MWEs badly.

MWEs can be classified into lexicalized phrases and institutionalized phrases which vary in their transparency and fixedness. In other words, generally MWEs can be classified into fixed type and unfixed type. However, the detailed standard of fixed MWEs is not given till now.

The foundamental core of an MWE can be understood in this way that the whole meaning of a series of word sequence can not be directly obtained from the single parts. According to Baldwin et al. (2003), MWEs can be classified into three categories concerning the compositionality: (1) non-compositional MWEs, where the meaning is opaque; (2) idiosyncratically compositional, where the words of component are unavailable outside the MWE in semantics; (3) simply compositional, where the word sequence is institutionalized. However, the classification can only work as an approximation, not as a concrete division. As Moon (1998) points out that compositionality can be regarded as a gradient along a continuum with no sharp demarcations, ranging from conventionalized, fully transparent literal expressions to completely opaque idiom.

In the literature, the fixedness of MWEs is set according to their compositionality and predictability. Many researchers were interested in noun compound, verb noun constructions, etc. For our purpose, a fixed MWE can be defined as a collocation of words that integrates closed and is used steadily.

However, we do not set the collection of part of speech. We use syntactic rules to filter collection of part of speech.

In this paper, we investigate experimentally the use of a multi-features frame for extracting fixed MWEs in domain-specific corpora. We put forwards the definition of fixed MWEs, calculate statistical measures including MI and entropy of a pos unit, filter MWEs candidates with the help of linguistics rules including syntactic rules, stop words. Finally, the fixed MWEs are obtained. This paper is organized as follows: Section II briefly discusses some previous works on methods for automatically extracting MWEs. Sections III presents the resources used in this paper. Section IV describes the approach to extract fixed MWEs. Section V presents the experiment and the result. Section VI finishes this paper with some conclusions and proposals for future work.

2 Related Works

Since 2003, a key workshop on MWEs has opened. A number of jobs has focused on the various aspects of MWEs, among which the identification and extraction are dominant. Due to the complex characters of MWEs, different approaches were adopted. In general, methods on the extraction of MWEs involve approaches: (1)

statistical methods (Piao et al., 2006; Zhang et al., 2006), (2) linguistic information (Baldwin et al, 2003; Bannard, 2007), (3) hybrid methods which combine the two approaches together(Baldwin and Vilavicencio, 2002; Cruys and Moirón, 2007).

A lot of energy has been put into the task of automatically extracting MWEs for English. A few papers focus on Chinese MWEs.(Duan et al., 2009, Zhixiang Ren et al., 2009). There is a new trend to extract Chinese MWEs.

There is no omnipotent approach to automatically extract MWEs. Statistical measures are effective to bigrams and trigrams. Many statistical measures are used for automatically extracting MWEs. However, it is still unclear that which statistical measure is the best one to carry out the task. Villavicencio et al.(2007) did experiment to compare statistical measures (mutual information, permutation entropy and $\chi 2$) for extraction of MWEs. The result is that Mutual Information seemed to differentiate MWEs from non- MWEs. Pearce (2002) evaluated statistical measures including Z score, Pointwise MI, cost reduction, left and right context entropy, odds ratio. The result is left and right context entropy is useful to determine the boundary of a unit.

However, statistical measures would bring much noise. Frequency is the core in some literature, which does not consider the syntactic and semantic information. In order to overcome the drawback, many statistical measures should work together.

In this paper, we put forward a hybrid method to extract MWEs for Chinese. Our approach is combined with statistical measures and linguistic rules and can be applied to other languages.

3 Data Resources

The texts used in the paper are from the Dynamic Circulating Corpus(DCC) which is composed of newspapers from 15 print media of main streams including People's Daily, Chinese Youth Daily, etc. Texts in the corpus are stored and filtered to remove ads. The size of DCC amounts to over 3 billion Chinese characters till now.

The POS software is offered by Institute of Automation, Chinese Academy of Science. The text classifier is offered by Institute of Applied Linguistics, Beijing Language and Culture University.

The machine-readable semantic dictionary named Hownet is also offered by Beijing Language and Culture University.

4 Extracting MWEs in the Multi-Feature Frame

The foundamental core of an MWE can be understood in this way that the whole meaning of a series of word sequence can not be directly obtained from the single parts. According to Baldwin et al.(2003), MWEs can be classified into three categories concerning the compositionality: (1) non-compositional MWEs, where the meaning is opaque; (2) idiosyncratically compositional, where the words of component are unavailable outside the MWE in semantics; (3) simply compositional, where the word sequence is institutionalized. However, the classification can only work as an approximation, not as a concrete division. As Moon(1998) points out that compositionality can be regarded as a gradient along a continuum with no sharp demarcations, ranging from conventionalized, fully transparent literal expressions to completely opaque idiom.

Since Chinese is an isolated language which is different from languages belonging to Indo-European language. Types of MWEs(such as Light verb constructions, Verb Particle Constructions) in western languages can not apply mechanically to Chinese. Hence, we do not set the part of speech in advance, which is different from western researchers who employ syntactic rules.

Among statistical measures, some measures(t-score, MI, x^2, Log-likelihood) can determine whether the inner combination of a word sequence is close. We regard such measures as inner measures. Meanwhile, some measures (left and right entropy) can be used to determine the independence of a word sequence, that is to say they can judge whether it is the boundary of a MWE. We regard such measures as outer measures.

Pecina(2010) points out that there is not a single universal measure to rank collocation, and suggests that different measures produce different results for different tasks considering data, language, and the various types of MWE after evaluating 82 lexical association measures to the rank the collocation candidates.

To extract effectively MWEs in our large-scale corpus, we choose statistical measures of frequency, MI, to detect whether a word sequence is close in inner part and entropy to detect whether a boundary of a word sequence is independent. Then, we employ linguistic rules and stop words to filter MWEs candidates, which can filter MWEs candidates not legal in linguistics.

In general, our approach is a hybrid method to combine statistical measures and linguistic rules. See the procedure:

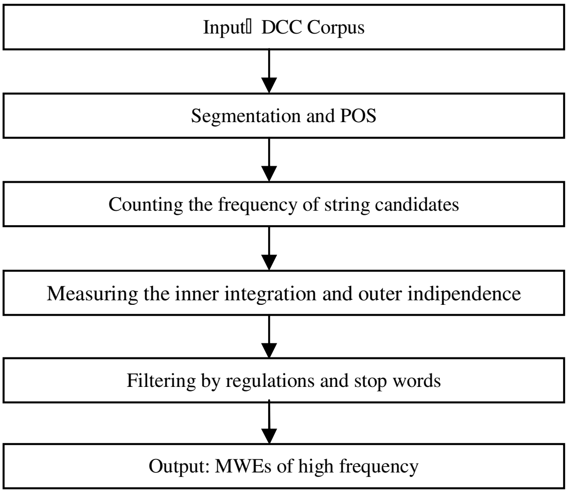

Input DCC Corpus
Segmentation and POS
Counting the frequency of string candidates
Measuring the inner integration and outer indipendence
Filtering by regulations and stop words
Output: MWEs of high frequency

Fig. 1. Digram of Procedure

With the help of statistical measures and syntactic rules, this paper judges the candidate MWEs from the perspectives of frequency, MI measures, left and right context entropy measures, syntactic rule to determine whether a MWEs is integrated closely; and judges whether candidate MWEs are legal in semantics. By means of diachronic analysis candidate MWEs are evaluated in the perspective of steady usage. After such extraction, over six hundred fixed MWEs are our result.

5 Experiment and Result

5.1 Identification of Seed Words in Special Domain

In order to build MWEs resource for Teaching Chinese as a Foreign Language and evaluate the result of MWEs, we select the special domain of education from texts of DCC in 2007. The size is 154M. First, we carry out text classification and part of speech.

Words with high frequency in a domain are likely to represent some character of the domain. In the texts we choose, words such as "学生、学校" show the character of the education domain. However, the top five words with the highest frequency are "的、是、在、一、了" can not represent the character of education domain. In order to solve this problem, we use the method of ID Comparison to choose words with domain characters.

ID Comparison method is to compare different ID of a word in different domains. The formula is as follows:

$$R_I = \frac{L(a,I)}{L(b,I)} \tag{1}$$

I represents one word; a and b represent different word table, L(a,I) represents the sequence of word I in table a, L(b,I) represents the sequence of word I in table b, R_I represents the ratio of word I between table a and table b.

For example, the word "教育" hold the 11st place in the domain of education. After ID Comparison, ID of the word "教育" becomes the first.

5.2 Extraction of Candidate Strings

After seed words are fixed, we open windows with such seed words and emerge them. In this step, we find bigram strings occupy 65% in the whole strings. Then we use statistical measures to filter them.

5.3 Filtering from Statistical Perspective

A large number of candidate strings produced from seed words are without syntactic regulation or semantic meaning. We employ statistical measures to filter such candidate strings. Statistical measures can both detect whether a candidate string is closely integrated inside and has a clear boundary outside. We choose mutual information and entropy to calculate and set the threshold.

5.4 Methodology to Judge the Internal Integration

From statistical perspective, whether a string is closely integrated depends on the co-occurance between words. If a string occurs repeatedly, it shows that the integration is strong. Therefore, a string with high frequency is likely to a MWE.

The formula is as follows:

$$\mathrm{MI}(a,b) = \log_2 \frac{P(a,b)}{P(a)P(b)} \tag{2}$$

The higher the MI is, the higher the integration between word a and word b is. In other words, word a and word b can make up a phrase to great extent. The lower the MI is, the lower the possibility of phrase between word a and word b is.

5.5 Methodology to Judge the External Boundary

In order to make a MWEs acceptable in semantic perspective, it must be an unabroken linguistic unit. Hung et al(2009). points out that left and right entropy can detect the composition of a string. Left entropy and Right entropy refer to the entropy of left boundary and right boundary of a MWE.

The formula is as follows:

$$Le(W) = -\sum_{\forall a \in A} P(aW \mid W) \cdot \log_2 P(aW \mid W) \tag{3}$$

$$Re(W) = -\sum_{\forall b \in B} P(Wb \mid W) \cdot \log_2 P(Wb \mid W) \tag{4}$$

Le and Re represent Left Entropy and Right Entropy of a string respectively; W represents a string of N-gram, $W=\{w_1,w_2 \ldots w_n\}$; A represents all words on the left of the string, a represents a word on the left, B represents all words on the right of the string, b represents a word on the right. The higher Le and Re, the more likely W is an decompositional MWE.

5.6 Filtering from Linguistic Perspective

To overcome the drawback of statistical measures, we use linguistic rules to filter continuously.

Stop words are words that can not appear at the beginning or at the end of a string. In such strings as “把家长、是孩子”, “是、把” are selected as stop words. The list of stop words are made annually.

By part of speech and stop words, we can filter more noise.

The algorithm is as follows:

```
Begin
Step 1. Input a string, if there is no stopword at the beginning and at the end, go to step 3, else delete the string and go to step 2.
Step 2. input next string.
Step 3. output the string.
Step 4. If input the last string, quit, else go to step 1, input next string.
```

The word formation of Chinese determines there is at least one headword which lies at the left or at the right of a MWE. By observation, we find that some word categories can not appear at the very beginning or at the very end of a MWE. In detail, particles and quantifiers can not appear at the very beginning; adverbs and conjunctions can not appear at the very end. Considering the characters of texts and such syntactic rules, we use linguistic information mentioned above to filter noise.

The algorithm is as follows:

```
Begin
Step 1. Input a string, if there is no limitation of particles at the beginning and at the end, go to step 3, else delete the string and go to step 2.
Step 2. input next string.
```

Step 3. output the string.
Step 4. If input the last string, quit, else go to step 1, input next string.

According to statistics, candidates with frequency over 50 occupy 8.33% in the whole dataset. We think they are prominent. Of course, threshold of frequency can be changed to meet the need of special purpose. Mutual information reflects the degree of correlation of two characters, if MI≤0, then there is no correlation between two characters. Hence, we choose MI>0 after calculation. By observation, we choose left entropy>3 and right entropy>3.

5.7 Results and Analysis

By annual analysis, the result shows that our method is efficient. Among all data, the precision is 98%. If the thresholds are changed, the result would also change. Hence, we can extract what we want according to special task and special purpose.

6 Conclusion and Future Works

This paper presents a hybrid method to extract rapidly MWEs. Successful extraction from large-scale corpus shows our method and computational efficiency are practicable. Our result was evaluated by human judgment, the result is considerable acceptable.

There are many directions to pursue in the future: (1) other measures should be calculated to find the best combination of association measures; (2) more combination of POS should be found and analyze; (3) Fixed MWEs larger than bigram should be examined.

Acknowledgments. This research has been funded by Taiyuan University of Technology (Item number: 900103-03010255 and 900103-03020632).

References

1. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002)
2. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: Grammar of Spoken and Written English. Longman, Harlow (1999)
3. Jackendoff, R.: The Architecture of the Language Faculty, Cambridge (1997)
4. Baldwin, T., Bender, E.M., Flickinger, D., Kim, A., Oepen, S.: Road-testing the English Resource Grammar over the British National Corpus. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, pp. 2047–2050 (2004)
5. Caseli, H.M., Ramisch, C., Nunes, M.G.V., Villavicencio, A.: Alignment-based extraction of multiword expressions. Language Resources and Evaluation (2009) (to appear)
6. Moon, R.: Fixed Expressions and Idioms in English: A Corpus-Based Approach. Clarendom Press, Oxford (1998)

7. Piao, S.S.L., Sun, G., Rayson, P., Yuan, Q.: Automatic Extraction of Chinese Multiword Expressions with a Statistical Tool. In: Proceedings of the Workshop on Multiword expressions in a Multilingual Context (EACL 2006), Trento, Italy, pp. 17–24 (April 2006)
8. Zhang, Y., Kordoni, V., Villavicencio, A., Idiart, M.: Automated Multiword Expression Prediction for Grammar Engineering. In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, pp. 36–44. Association for Computational Linguistics, Sydney (July 2006)
9. Bannard, C.: A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In: Proceedings of the ACL Workshop on A Broader Perspective on Multiword Expressions, pp. 1–8 (2007)
10. Baldwin, T., Villavicencio, A.: Extracting the Unextractable: A Case Study on Verb-particles. In: Proceedings of the 6th Conference on Natural Language Learning (CoNLL 2002), Taipei, Taiwan, pp. 98–104 (2002)
11. Van de Cruys, T., Moirón, B.V.: Semantics-based multiword expression extraction. In: Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, pp. 25–32 (2007)
12. Duan, J., Zhang, M., Tong, L., Guo, F.: A Hybrid Approach to Improve Bilingual Multiword Expression Extraction. In: Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 541–547. Springer, Heidelberg (2009)
13. Ren, Z., Lu, Y., Cao, J., Liu, Q., Huang, Y.: Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pp. 47–54. Suntec, Singapore (2009)
14. Villavicencio, A., Kordoni, V., Zhang, Y., MarcoIdiart, Ramisch, C.: Validation and Evaluation of Automatically, Acquired Multiword Expressions for Grammar Engineering. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June 2007, pp. 1034–1043 (2007)
15. Pearce, D.: A comparative evaluation of collocation extraction techniques. In: Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, pp. 1530–1536 (2002)
16. Pecina, P.: Lexical association measures and collocation extraction. Language Resources and Evaluation 44, 137–158 (2010)
17. Hoang, H.H., Kim, S.N., Kan, M.-Y.: A Re-examination of Lexical Association Measures. In: Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, Suntec, Singapore, pp. 31–39 (2009)
18. Davidov, D., Rappoport, A.: Efficient Unsupervised Discovery of Word Categories Using Symmetric Patterns and High Frequency words. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, pp. 297–304 (July 2006)
19. Jackendoff, R.: The Architecture of the Language Faculty, Cambridge (1997)

Mining Popular Menu Items of a Restaurant from Web Reviews[*]

Yeong Hyeon Gu and Seong Joon Yoo[**]

Department of Computer Engineering, Sejong University,
98 Gunja-Dong, Gwangjin-Gu, Seoul, Korea
sjyoo@sejong.ac.kr

Abstract. We propose a novel method to mine popular menu items from online reviews. In order to extract popular menu items, a crawler that uses the wrapper on search web sites was used to collect online reviews, restaurant names, and menu items. Then, unnecessary posts were removed by using the patterns. Also, post frequency was used to find the most frequently appearing menu items from online reviews in order to select the most popular menu items. In the result, the total average accuracy was 0.900.

Keywords: menu item extraction, popular menu item selection.

1 Introduction

In the past, people acquired information for good restaurants from friends or other close people. However, these days, people acquire good restaurant information from the internet and smart phones. In order to select the most popular menu items at a restaurant, it needs more efforts. For example, if the name of a restaurant includes 'Food A'. Most people will assume that the most popular food is 'Food A', so they will order that one. However, sometimes, 'Food B' is more popular and more delicious than 'Food A'.

Most restaurant information services don't provide the popular menu items, and if they do, it is only for a limited number of restaurants. Therefore, people need to search blogs, and communities on the internet to find other people's reviews on menu items. Because they have to search many posts on the web and need to read every single review, it consumes time and effort.

Moreover, current restaurant information services can be searched only with restaurant names or category information, and it is impossible to search the restaurants that provide certain kinds of menu items. In this case, the administrator needs to manually enter the menu items by each restaurant. Because the administrator needs to manually enter the menu items, it consumes lots of time and efforts to select the popular menu items. Therefore, restaurant information services are unable to

[*] This work was supported by the Korea Research Foundation(KRF) grant funded by the Korea government(MEST) (No. 2011-0002899).

[**] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 242–250, 2011.

provide service about the entire restaurants but only provide restaurants in a certain scope that can be handled manually.

Moreover, existing menu item recommendation methods reflected the website operator's personal favorite menu item or the menu items that were being promoted by advertisers. Therefore, actually, the recommendation based on actual experiences without being engaged to companies is the most trustworthy information.

In order to solve such problems, the methods to extract the menu items from online reviews and to select popular menu items based on the extracted menu items were suggested in this paper.

In order to extract popular menu items, a crawler using the wrapper [1][2][3][4] on search web sites and websites was used to collect online reviews, restaurant names, and food menu items. There are many unnecessary review posts that decrease the accuracy of popular menu item selection, and these are removed using the pattern rules.

There is a high possibility that the menu items that frequently appeared in online reviews for restaurants are the menu items from that restaurant. Also the menu items that are mentioned most frequently may be the most popular menu item of that restaurant. Therefore, online reviews for certain restaurants were collected and the menu items that appeared most frequently were extracted. Then, the menu items that appeared most frequently were selected as the popular menu items.

In Chapter 2 of this paper, related works are analyzed. In Chapter 3, online restaurant review post collection is described. In Chapter 4, selection and extraction methods of menu items are described and in Chapter 5, the popular menu item selection methods are described. In Chapter 6, the performance evaluation results are described through experiments. Chapter 7 is the conclusion.

2 Related Works

In order to select popular menu items from online reviews, the crawler needs to be used on existing searching websites and blogs to collect the online reviews, restaurant names, and menu items.

Generally web crawlers go to web servers to analyze webpage contents and extract the included URLs. Then, it moves to the next URL to collect web posts. The various and high volume of web posts collected are used by the search engine.

Meanwhile, there are crawlers that can only collect a certain subject of posts. Focused Crawler [5] and Topical Crawler [6][7] are equipped with post classifiers or rules of posts to be collected. However, if the classifier's performance is poor and the volume of rule is insufficient, it can't collect enough of the data that users want.

Due to such reasons, a research on wrappers that can analyze the website structure to directly collect data has been performed. Crawlers that use wrappers establish rules in order to extract needing data and automatically collect the contents of websites. However, such contents contain high volume of unnecessary information which decreases the accuracy. Therefore, in this paper, the rules were used again for re-classification.

In order to find popular menu items, restaurant names and menu items need to be collected from the posts that have been collected through crawler. When extracting information from unstructured text, information such as restaurant name and menu items are very important. Such information is called a ‘named entity’, and recognizing the named entity from text is called ‘named entity recognition [8]’.

Studies on ‘named entity recognition’ have been conducted not only in English but in various languages, and been utilized in various domains. Usually the studies were regarding the proper nouns, time, and number recognition. However, these days, studies on movies, email addresses, phone number, books, jobs, product brand names, and bioinformatics recognition have been conducted.

In addition, there are also studies on named entity recognition for open domains [9]. In [10], newspaper corpus, existing systems, and thesaurus were used to create the rules and 200 named entity categories were classified.

Since there are high volumes of products and service related reviews on the web, many opinion mining studies have been actively conducted. Opinion mining [11][12][13][14][15] finds effective information from high volume review data. The representative use of the opinion mining is to extract product reviews.

In [14], the product related characteristics were extracted by means of the rules that were created based on the number of frequencies from users’ reviews. Then, reviews were classified into positive and negative review posts by each extract characteristic.

[15] collected comments from shopping mall purchasers. Then reviews were classified into positive and negative review posts by each extract characteristic to summarize user’s opinions.

[16] found comparative sentences from online reviews and analyzed the components of the sentences to extract the comparative targets and characteristics. Then, the targets that are on comparative advantage were extracted. For this, comparative sentences were divided into 4 types and appropriate rules for each type were applied to extract the targets that have comparative advantage.

Various studies have been conducted to collect online reviews and to acquire necessary information. However, there have been no studies extracting menu items and selecting popular menu items based on the extracted menu items. Therefore, in this work, menu items were extracted from online reviews and popular menu items were selected based on the extracted menu items.

3 Collecting the Online Restaurant Reviews

It is difficult to collect online reviews for restaurants from the high volume of posts online. Therefore, a somewhat roundabout method was used in this work. First of all, restaurant names were collected by means of web crawler from restaurant searching websites. The way to accurately collect is using html code of the website to analyze the website structure. Then, the access should be made only to the areas where the data is. The program that collects the data based on the website structure information is the wrapper. In this paper, the restaurant names and menu items were collected from the restaurant searching websites based on the wrapper model.

Then, restaurant names were given as a query to a portal searching engine then, the web crawler was used again to collect the results. There are various types of online restaurant reviews however, in this work it is limited to blog posts.

In the result of searching, many posts that were irreverent to restaurants were collected even though restaurant names were given as query. In order to solve such a problem, phone numbers of the restaurants in addition to the names were collected from restaurant searching websites. Both restaurant names and phone numbers were used as a query.

After collecting the online reviews, the listing type of restaurant posts were removed from the collected posts. Since the posts that listed good restaurants include various menu items and restaurants, it highly decreases the extracted menu item results quality. Examples are ‘good restaurants near A station’, ‘Good restaurants list in B area’, and ‘Good restaurants that were on TV’. These posts include recommended menu item of restaurant_1, recommended menu item of restaurant_2, …, restaurant_n. Therefore, it is difficult to determine the recommended menu item of the restaurant before dividing the contents to each segment. For this reason, listings were excluded in this research.

Various restaurant listings have common elements. Most listings include the restaurant phone numbers. Since more than one restaurant is introduced, it includes more than two phone numbers. Therefore, the posts that include more than two phone numbers were excluded in order to exclude restaurant listings.

4 Menu Item Selection and Extraction

In order to extract menu items from online restaurant review posts, a menu items dictionary were established in advance. Menu items were extracted among the words from the restaurant review posts by using the pre-established menu item dictionary. It takes a lot of time to manually establish the menu item dictionary. Therefore, restaurant search engines [17][18][19][20][21] were used to collect the menu items by using a web crawler. Total 34,506 of menu items were collected without duplication, and these menu items were sorted by the frequency. It is because the frequently appeared menu items are more popular and common menu items. Top 1,000 menu items were selected among the collected menu items. The reason of selecting the menu items with high frequency was because the menu items with low frequency were not common menu items for most restaurants. For example, let’s say that ‘Menu item1’ is called ‘A restaurant’s Menu item 1’ at A restaurant. Even though these are the same food, because ‘Menu item 1’ is a commonly called name, the frequency is high, and the frequency of ‘A restaurant’s Menu item 1’ is low because the name is used only at A restaurant.

Among the top 1,000 menu items, some menu items were beverages or alcohols which don’t qualify the study criteria. Therefore unqualified menu items were removed and remaining were registered to the menu item dictionary.

In order to extract menu items from the online restaurant review posts that were collected by using the web crawler, first morpheme analysis[22] was conducted.

Then, only the nouns were selected to compare with the pre-established restaurant menu item dictionary. At this time, the matching nouns were extracted from titles, tags, and body text. The extracted menu items and frequencies were saved in DB.

5 Popular Menu Item Selection

In the previous restaurant information service, administrators manually entered the menu items of restaurants. Although Wingspoon[17] manually provides popular menu item information, most restaurants information websites don't provide popular menu item information. It is because it takes too much efforts and time to manually enter the menu items and choose popular menu items.

In order to solve such problem, frequencies were used to choose popular menu items of each restaurant. There are many online reviews on evaluations that people made after visiting the restaurant. Restaurant review posts not only evaluate restaurants but also mention and evaluate the menu items that were provided by that restaurant. Therefore, the menu items that often appeared from online reviews collected for a certain restaurant are most likely the menu items provided by that restaurant. Moreover, the menu items that are mentioned the most are most likely the most popular menu item of that restaurant. Therefore, the online reviews on certain restaurants were collected and the menu items that appeared most frequently were extracted. And these menu items were selected as the most popular menu items of that restaurant.

At that time, if the term frequency is used, there is a high possibility to induce incorrect results due to certain restaurant review posts. For example, let's say that there are three review posts on restaurant A and the frequencies of each menu item are as in Table 1.

Table 1. Example of using term frequency

Division	Frequency of Menu item 1	Frequency of Menu item 2	Frequency of Menu item 3
Post 1	8	1	0
Post 2	0	2	4
Post 3	1	2	3

Even though Menu item 3 had the most frequency among 2/3 of the entire posts, menu item 1 shows the highest frequency 9 in the aspect of the sum of the entire frequency. It is because post 1 mentioned menu item 1 too many times. Therefore, if term frequency is used to calculate the frequency, incorrect results can be induced.

To avoid such a problem, post frequency rather than term frequency was used. Therefore, the number of review posts that included certain menu items was counted.

Fig. 1 shows the system structure extracting menu items from online reviews and selecting the popular menu items.

```
// load data
menu_db = loadDB(type_menu)
blog_data=blogCrawling()

// data processing
result = preprocessing(blog_data)
pos_result = POSTagging(result)
menuList = menuExtracting(pos_result, menu_db)
menu_struct = menuFrequencyCounting(menuList)
menu_struct = sort(menu_struct , freq_asc)
```

loadDB (): load the pre-established restaurant menu item dictionary
blogCrawling():collect the online restaurant review posts
preprocessing():exclude review include more than two phone numbers
POSTagging(): tag the part of speech
menuExtracting(): extract the matching nouns from titles, tags, and body text
menuFrequencyCounting(): count the frequency of menu
sort():menu items are sorted by the document frequency

Fig. 1. Algorithm of popular menu items selection

6 Experiment

In this chapter, the performance evaluation results are described through experiments.

6.1 Experiment Data

For the experiment, blog posts were collected from various web posts. 42,126 Restaurant review blog posts were collected by using crawler. Among them, there were 24,882 review posts after excluding restaurant listing type of posts. There were 14,803 review blog posts that included menu items, and there were 39,990 menu items extracted from restaurant review blogs. Up to 5 menu items were extracted from each blog to prevent the extraction of too many menu items.

Among the extracted menu items, 7,620 were extracted from blog titles, 7,049 were extracted from blog tags, and 25,321 were extracted from blog body text. A total of 3,184 restaurants were used for the experiment and a total of 14,507 popular menu items were extracted from the restaurant review blogs.

6.2 Performance Evaluation

In order to evaluate the extracted popular menu items from the restaurant review blogs, those were compared with the actual restaurant menu items.

The accuracy of the popular menu items extracted from review post titles was 0.885 and 0.936 from tags, 0.878 from body texts, and the total accuracy was 0.900. The accuracy of the menu item extracted from tags was the highest and the accuracy extracted from body text was the lowest.

Many menu items of which post frequency was only 1~3 were not actual menu items from the restaurants. Therefore, only 5 menu items per restaurant were selected as the popular menu item to remove less frequently appeared menu items. However, since most restaurants have more than 5 menu items, reproduction ratio was not acquired and only the accuracy was measured. Accuracy of Popular Menu items.

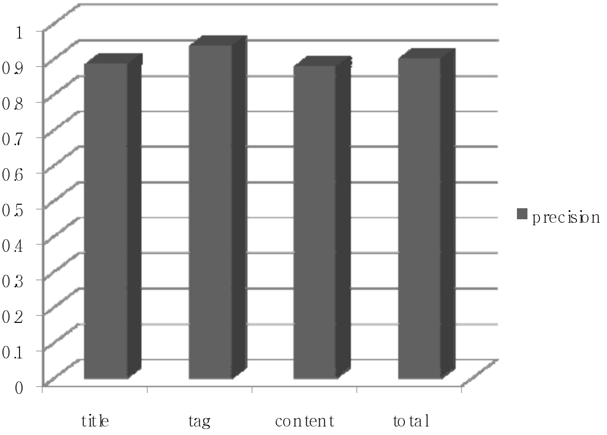


Fig. 2. Accuracy of popular menu items

In many cases in which incorrect popular menu items got selected, the crawling of restaurant review posts were incorrect. Therefore, in order to increase the performance of selecting popular menu items, the most important thing is improving the web crawler. Also phone number pattern was used to remove restaurant listing type of posts, however, the listing type of posts that didn't have that exact pattern were not removed. For example, if there is a space after the hyphen ('777- 2269'), this is out of the phone number pattern. Therefore, it is necessary to make a rule that can cover more cases by assuming more various cases.

7 Conclusion

In this study, methods to extract menu items from online reviews and select popular menu items based on the extracted menu items were suggested.

Existing menu item recommendations reflected the website operator's favors or advertiser's promotions. Therefore, actually, the recommendation based on actual experiences without being engaged to companies is the most trustworthy information. So, the popular menu items that are indirectly recommended by consumers can be found through the menu items that are most frequently mentioned from online restaurant reviews.

In order to extract popular menu items, restaurant names and menu items were collected from existing restaurant search engines by using a crawler that uses a wrapper. Then, unnecessary posts were removed by using patterns. Then, post frequency was used to select the most frequently appeared menu items on the online reviews as the most popular menu items.

Review posts were classified into title, tag, and body text to select the popular menu items. In the experiment result, the popular menu items from tag showed the highest accuracy and the entire average accuracy was 0.900.

In this time, the online reviews that contain both restaurant names and phone numbers were collected. And many online reviews that only include restaurant names were not collected. Therefore the accuracy is poor for some restaurants due to the small number of collected online reviews or some only reflected the tastes of a small number of reviewers. In order to improve this issue, the online reviews that include only the restaurant names will be collected for future studies, which will increase the whole volume of review posts as well as the accuracy.

Also, only blog posts were targeted for this study; however, regular web posts and social network posts also need to be considered in the future.

References

1. Chang, C.H., Kayed, M., Girgis, M.R., Shaalan, K.: A Survey of Web Information Extraction Systems. IEEE Transaction on Knowledge and Data Engineering 18(10), 1411–1428 (2006)
2. Bertoli, C., Crescenzi, Y., Merialdo, P.: Crawling programs for wrapper-based applications. In: IEEE IRI 2008, pp. 160–165 (2008)
3. Yang, J., Kim, T., Choi, J.: An Interface Agent for Wrapper-Based Information Extraction. In: Barley, M.W., Kasabov, N. (eds.) PRIMA 2004. LNCS (LNAI), vol. 3371, pp. 291–302. Springer, Heidelberg (2005)
4. Soderland, S., Cardie, C., Mooney, R.: Learning information extraction rules for semi-structured and free text. Machine Learning (1999)
5. Chakrabarti, S., Berg, M., Dom, B.: Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. Computer Networks 31(11-16), 1623–1640 (1999)
6. Chakrabarti, S.: Mining the Web, Discovering Knowledge from Hypertext Data. Morgan Kaufmann, San Francisco (2003)
7. Cho, J., Garcia, H., Page, L.: Efficient Crawling through URL Ordering. Computer Networks 30(1-7), 161–172 (1998)
8. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
9. Alfonseca, E., Manandhar, S.: An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. In: International Conference on General WordNet, pp. 1–9 (2002)
10. Satoshi, S., Nobata, C.: Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In: Conference on Language Resources and Evaluation, pp. 1977–1988 (2004)
11. Liu, B.: Web Data Mining. Springer, Heidelberg
12. Tuerny, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment Classification using Machine Learning Techniques. In: Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)

14. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
15. Hu, M., Liu, B.: Mining Opinion Features in Customer Reviews. In: 19th National Conference on Artificial Intelligence(AAAI 2004), pp. 755–760 (2004)
16. Jindal, N., Liu, B.: Mining Comparative Sentences and Relations. In: AAAI 2006 (2006)
17. Wing Spoon, http://www.wingspoon.cokr
18. Menupandotcom, http://www.menupan.com
19. Local Story, http://www.localstory.kr
20. Food N Cafe, http://www.foodncafe.com
21. Daum Place, http://place.daum.net
22. Gang, S.: Analysis of Korean Morphemes and Information Retrieval. Hungrung Publish (2002)

News Information Extraction Based on Adaptive Weighting Using Unsupervised Bayesian Algorithm

Shilin Huang[1], Xiaolin Zheng[1], Xiaowei Wang[2], and Deren Chen[1]

College of Computer Science and Technology, Zhejiang University
Hangzhou, China
{supercat,xlzheng,drchen}@zju.edu.cn,
raindreams@126.com

Abstract. Information extraction is important in web information retrieval. In case of news information extraction, because news information does not have representative keywords pointing out its beginning and ending, it is difficult to specify the news title and body automatically. Our approach is based on an adaptive weighting factor using Bayesian algorithm to solve this problem. We divided a news page into text fragments, and represented them with a set of content features and layout features. We used an adaptive weighting factor to make features fit in different pages. Experiments show that our method results in a higher precision than the original algorithm without a weighting factor on the task of news information extraction.

Keywords: information extraction, adaptive weighting.

1 Introduction

Internet is becoming a more and more important source of information. Information extraction is to automatically extract factual information from web pages and convert it into well-structured format for later analysis. We studied news information extraction in this paper. Base on our result, we can later extract factual information such as the time and place of an event, or product information from newly present finance news. Also, the screen reader users and mobile device users will feel more convenient with the extracted news information from our work.

Our approach is based on adaptive weighting factor using unsupervised Bayesian algorithm. It can efficiently extract the news title and content. We divide pages into text fragments and represent them with a set of layout and content features. Considering that each feature has different ability on describing different web pages, we add adaptive weighting factors to the features. Experiment shows a good performance of our approach.

2 Related Works

Early approaches like wrapper induction systems [7, 14] and Roadrunner [5] retrieved templates by analyzing page similarity. They could not handle when new templates

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 251–258, 2011.

occurred. MDR [10] identified the similarity of page blocks with their edit distance, but it made mistakes when noise was interposed. Zhai, Y. and B. Liu [27] developed a method based on partial tree matching. Simon, K. and G. Lausen [18] introduced extra information to extend MDR. Zhai, Y. and B. Liu [26] proposed a method based on instance study and it could learn a new template when new instance came up. Zhao, H., W. Meng, and C. Yu [28] recorded the templates by a graph model. Shoubiao, T., F. Jin, and J. Yuan [16] eliminated the conception ambiguity basing on a label library, and used MDR to find the repeating mode. Shuyi, Z., et al [17] proposed an algorithm based on the distance between the page and the wrapper. Tak-Lam, W. and L. Wai [19] developed a wrapper to learn features from the source site which can apply on new sites through modeling the dependency between different web sites.

Many researches were based on machine learning and statistical learning methods. Miao, G., et al. [12] recognized the similarity of two label paths through the comparison of occurrence mode of the label paths. Carlson, A., et al. [3] used a semi-supervised method to train wrappers for different class relationship. Approach of Xiao Jian-Peng, Z.L.-S. and Ren Xing [25] was based on TSVM. It only classified new data vectors related and needed very few labeled sample. Labský, M., et al. [8] tried to develop a tool based on well-structured ontology to construct the model semi-automatically. Junfang, S. and L. Li [6] developed a news extraction method based on news domain ontology. Wong, T.-L. and W. Lam [23] modeled the dependency between text blocks from one or several pages with an undirected graph model. Michal Mared, P.P., Miroslav Spousta [13] classified page fragments with conditional random field algorithm. Pasternack, J. and D. Roth [15] developed a maximum sub-string division method to find the fragment containing the article.

A different kind of methods analyzes visual features of the pages. VIPS [1, 2] was a classic unsupervised algorithm. It could extract the text precisely but it consumed a lot of system resource. Ma, L., et al. [11] chose to divide the page by the <TABLE> tag and Lin, S.-H. and J.-M. Ho [9] introduced the measure of entropy based on word features to make division more precisely. Chen, L., S. Ye, and X. Li [4] clustered fragments by similar style and position. Vadrevu, S., F. Gelgi, and H. Davulcu [20] organized the pages like XML and introduce domain knowledge to build a statistical model. Wang, J., et al. [21, 22] learn the horizontal and vertical relationship of positions from a few training samples, and extract the news with its spatial features.

3 Unsupervised Bayesian Method with Adaptive Weighting Factor

3.1 Unsupervised Bayesian Algorithm

We consider a web page as a sequence of text fragments, and can be represented by a set of features, both content and layout features. Depending on the characteristic of job information, Wong, T.-L., W. Lam, and B. Chen [24] proposed an unsupervised method based on the Naïve Bayesian Theory to discover headings of job information. They used variable h to represent whether a text fragment f is a heading or not. Assume that f can be represented by a set of features $\mathrm{A} = \{\mathrm{a}_1, \mathrm{a}_2, \ldots, \mathrm{a}_\mathrm{n}\}$. The classic Bayesian algorithm gives out a method to compute the probability of f to be a heading $\mathrm{P(h|f)}$ basing on the prior probability $\mathrm{P(h)}$. By observing $\mathrm{P(f|h)}$ in given case, we can learn that, $P(h|f) = \frac{P(f|h)P(h)}{P(f)}$.The Naïve Bayesian Theory is based on a supposition that all a_i are exclusive. So we know that $P(h|f) \propto P(h)\prod_{i=1}^{n} P(a_i|h)$.

As different web pages have different layouts, an iterative algorithm based on the Expectation Maximization (EM) Algorithm was introduced in [24],

E-STEP

$$P^{t+1}(h|f) = P^t(h)\prod_{i=1}^{|A|} P^t(a_i|h), \forall f \in F^i \tag{1}$$

M-STEP

$$P^{t+1}(A_i(\cdot) = 1|h) = \frac{\sum_{f \in F^i} A_i(f)P^{t+1}(h|f)+\delta}{\sum_{h\prime=1,0}\{\sum_{f \in F^i} A_i(f)P^{t+1}(h\prime|f)+\delta\}} \tag{2}$$

$$P^{t+1}(h) = \frac{\sum_{f \in F^i} P^{t+1}(h|f)+\delta}{|F^i|+2\delta} \tag{3}$$

Where F^i represents the text fragments of the i-th page, δ is the smoothing factor to avoid zero probability. In E-STEP, $P^{t+1}(h|f)$ in the (t+1)-th iteration of all $f \in F^i$ is computed, and then in M-STEP $P^{t+1}(A_i(\cdot)|h)$ and $P^{t+1}(h)$ are updated.

3.2 Algorithm with Adaptive Weighting Factor

This algorithm performed well on job information extraction, but in news information, a feature may have different abilities on describing different pages. For example, word “company” may be important in business related pages, but normal in other pages. We introduced a weighting factor to deal with this problem. After division of a page, the task of information extraction is actually the classification of text fragments. The feature weight is the ability of a feature classifying different categories.

TF-IDF is a normal weighting technology. TF represents term frequency, which is usually defined as $tf(t,d) = \frac{n_{t,d}}{\sum_k n_{t,k}}$. $n_{t,d}$ is the occurrence number, and $\sum_k n_{t,k}$ is the normalization factor.IDF represents the inverse document frequency. It can be obtain as $idf(t) = \log\frac{|D|}{|\{d:d\ni t\}|+1}$. Here, |D| means the amount of documents, and $|\{d: d \ni t\}|$ means the amount of documents containing word t. However, TF-IDF only considers the occurring frequency of a word and its distribution in a page. This makes it more likely to classify a fragment to a larger category. Because there is much noise in a page, it could result in interruption. Our weighting factor would adjust adaptively. If a feature is more likely to be owned by a target, we increase its weight, and otherwise we decrease it. We modify the algorithm in [24] as follow,

E-STEP

$$P^{k+1}(t|f) = P^k(t) \times \prod_{m=1}^{|L|}(P^k(L_m(f)|t) \times W^k(L_m)) \times \prod_{n=1}^{|C|}(P^k(C_n(f)|t) \times W^k(C_n)), \forall f \in F^i \tag{4}$$

M-STEP

$$P^{k+1}(C_n(\cdot) = 1|t) = \frac{\sum_{f \in F^i} C_n(f)P^{k+1}(t|f)+\delta}{\sum_{t\prime=1,0}\{\sum_{f \in F^i} C_n(f)P^{k+1}(t\prime|f)+\delta\}} \tag{5}$$

$$\mathrm{P^{k+1}(L_m(\cdot) = 1|t) = \frac{\sum_{f\in F^i} L_m(f)P^{k+1}(t|f)+\delta}{\sum_{t'=1,0}\{\sum_{f\in F^i} L_m(f)P^{k+1}(t'|f)+\delta\}}} \tag{6}$$

$$\mathrm{P^{k+1}(t) = \frac{\sum_{f\in F^i} P^{k+1}(t|f)+\delta}{|F^i|+2\delta}} \tag{7}$$

$$\mathrm{W^{k+1}(L_m) = W^k(L_m) \times \frac{P^{k+1}(L_m(\cdot)=1|t)}{P^k(L_m(\cdot)=1|t)}} \tag{8}$$

$$\mathrm{W^{k+1}(C_n) = W^k(C_n) \times \frac{P^{k+1}(L_m(\cdot)=1|t)}{P^k(L_m(\cdot)=1|t)}} \tag{9}$$

Here, $\mathrm{F^i}$ represents the text fragments of the i-th page, and δ is a smoothing factor avoiding 0 probability. In E-STEP, we compute $\mathrm{P^{k+1}(t|f)}$ for all $\mathrm{f \in F^i}$ according to the result in the k-th iteration. In M-STEP, according to $\mathrm{P^{k+1}(t|f)}$ learnt in the E-STEP, we update $\mathrm{P^{k+1}(C_n(\cdot)|t)}$ and $\mathrm{P^{k+1}(L_m(\cdot) = 1|t)}$ and $\mathrm{P^{k+1}(t)}$. $\mathrm{W^k(L_m)}$ and $\mathrm{W^k(C_n)}$ represent the weight of the layout feature $\mathrm{L_m}$ and the weight of the content feature $\mathrm{C_n}$ in the k-th iteration respectively. The iteration repeats until we get a best result or a specific number of loops are reached.

3.3 Preprocessing and Initialization

In this paper, we divide a page by the minimum text nodes. A minimum text node is the smallest sub-tree only containing text. If a node does not contain any script like nodes, we consider it and its children nodes to be a minimum text sub-tree. Else, we divide it by its children nodes and until the page is totally divided. The layout feature set L includes features like bold, italic, color, long-text, short-text, and etc. It was manually preset whereas the content feature set is learned with a feature extractor.

For document independent content features, we have learnt their weights during the feature extraction process. To avoid too small value in the later calculation, we added a smoothing factor ε. The initial weight of the content feature is as follow,

$$\mathrm{W^0(C_n) = tfidf(C_n) + \varepsilon.} \tag{10}$$

On the other hand, the layout feature is manually given and they are document dependent. We calculate their initial weights for every single page as follow,

$$\mathrm{W^0(L_m) = \frac{tf(L_m(f)=1,F^i)\times idf(L_m(f)=1)}{\sqrt{[\sum_{m=1}^{|L|} tf(L_m(f)=1,F^i)\times idf(L_m(f)=1)]^2}}} \tag{11}$$

Here, $\mathrm{tf(L_m(f) = 1, F^i) = \frac{n_{L_m(f)=1}}{\sum_{m=1}^{|L|} n_{L_m(f)=1}}}$, and $\mathrm{idf(L_m(f) = 1) = \log\frac{|F^i|}{n_{L_m(f)=1}}}$, where $\mathrm{tf(L_m(f) = 1, F^i)}$ represents the term frequency of the m-th feature $\mathrm{L_m}$ in $\mathrm{F^i}$, and $\mathrm{idf(L_m(f) = 1)}$ is the inverse document frequency, $\mathrm{n_{L_m(f)=1}}$ means the occurrence number of $\mathrm{L_m}$ and $\mathrm{|F^i|}$ means the total number of page fragments. The denominator of (14) is the normalization factor.

$\mathrm{P^0(t|f)}$ is also unknown. Wong, T.-L., W. Lam, and B. Chen [28] used a seed word set to do the initialization. However, it is more complicated in news information. We use a simple binary classifier to initialize. If the classifier classifies a fragment to be a target, we set $\mathrm{P^0(t|f)}$ to greater than 0.5 and otherwise smaller than 0.5.

After we get a best result or a specific number of loops are reached, we return the final $P^{k+1}(t = 1|f)$. For all fragments with probability$P^{k+1}(t = 1|f) > P^{k+1}(t = 0|f)$, we consider them to be probable output. News content may be contained in several text fragments, so we output all these fragments. But news title can only have one, and we consider the fragment which has the highest probability as the final output.

4 Experiments and Analysis

To verify the availability of the adaptive weighting factor, we designed several experiments. The test pages used in the experiments were crawled from three famous news websites, which were Yahoo! News, NY Times, and China Daily respectively. We sorted these pages into 8 categories, including society, entertainment, sports, education, technology, health, business and politics. We randomly choose several extracted news pages to construct the training sets of the feature extractor and the binary classifier. The two training sets are exclusive to each other, and also to the testing set. We manually labeled the correct news for all pages in the test set to contrast with our extracting result.

Table 1. Contrast of Accuracy with Separate Training Set and Unified Training Set

	Precision with Separate Training Set			Precision with Unified Training Set			Page #
	Title	Content	Total	Title	Content	Total	
Society	99.5%	99.3%	99.0%	100%	99.3%	99.3%	404
Entertainment	100%	99.1%	99.1%	100%	98.8%	98.8%	339
Sports	100%	99.1%	99.1%	100%	99.4%	99.4%	674
Education	99.4%	99.1%	98.7%	99.7%	99.1%	98.7%	319
Technology	99.7%	100%	99.7%	100%	99.7%	99.7%	339
Health	99.8%	99.8%	99.8%	99.8%	99.6%	99.4%	494
Business	99.1%	98.9%	98.6%	99.5%	98.6%	98.2%	435
Politics	99.7%	98.7%	98.7%	100%	99.7%	99.7%	378
Total	99.7%	99.2%	99.1%	99.8%	99.3%	99.2%	3382

We knew that the content feature is independent on the websites, but it should be dependent on the domain. We first choose 30 pages for each category to train separate content feature sets. Then we mix them together to train a unified content feature set. Table 1 is the contrast of the precision with separate training set and unified training set. The result shows that a unified training set do not reduce the accuracy. Although the unified trained features can be quite different from the ones trained separately, the adaptive weighting factor performed well to make a good result. With the adaptive weighting factor, the features can perform differently in different categories. Thus we don't need to train them separately and they can adjust by themselves.

Then we test the performance of our algorithm on different templates. We use a unified training set for the content feature and only observe how the layout feature affects the result. Table 2 contrasts the performance of the weighting factor on different web sites. The result shows that templates do not influence much. The iterative learning process of the probability of the layout features and their corresponding weighting factor performs well, and totally get accuracy at 99.2%.

Table 2. Contrast of Accuracy of Different Web Sites

	Title Precision	Content Precision	Total Precision	Page #
Yahoo! News	100%	99.7%	99.7%	1127
NY Times	99.9%	99.3%	99.1%	1242
China Daily	99.8%	98.9%	98.7%	1013
Total	99.9%	99.3%	99.2%	3382

To observe the advantage of the adaptive weighting factor, we make a contrast between our method and the original unsupervised Bayesian algorithm without a weighting factor on news information extraction. Table 3 is the contrast of the original algorithm and our approach with the weighting factor.

Table 3. Contrast of Original Unsupervised Method and Our Method with Weighting Factors

	Precision of the original unsupervised Bayesian algorithm			Precision of our method with the weighting factor			Page #
	Title	Content	Total	Title	Content	Total	
Society	97.5%	94.8%	94.6%	100%	99.3%	99.3%	404
entertainment	98.2%	97.4%	96.7%	100%	98.5%	98.5%	339
sports	95.2%	94.2%	93.8%	100%	99.3%	99.3%	674
education	91.2%	84.6%	84.3%	99.7%	99.1%	98.7%	319
technology	93.8%	93.5%	93.2%	100%	99.7%	99.7%	339
health	94.8%	92.1%	88.9%	99.8%	99.6%	99.4%	494
business	98.8%	97.7%	97.3%	99.5%	98.9%	98.4%	435
politics	92.3%	92.6%	90.5%	100%	99.7%	99.7%	378
Total	95.4%	93.6%	92.6%	99.9%	99.3%	99.1%	3382

The result shows that the adaptive weighting factor improves the extraction result obviously. Without the weighting factor, content features cannot well specify texts of different domains. In our approach, the weighting factor adaptively changed when it could not match with the probability of the content feature. The content features that are not related to the current page would have a smaller weight. In the case of layout feature, because features could have different importance in different websites, the layout feature weight would differ in different web pages. Hence, our method reaches a much higher accuracy than the original algorithm on the news information extraction task. This is also the reason why the result in table 1 does not diverse a lot.

In our experiments we also found that the mistaken news contents are usually a missing of text fragments among continual content fragments. However, noises are seldom misjudged to be targets. We consider that it is because granularity we chose was small enough to avoid noises. But we did not consider the relationship among text fragments, so we sometimes missed fragments not continually.

5 Conclusions

According to the characteristics of news information, we improve the unsupervised algorithm based on naïve Bayesian algorithm to fit in news information extraction task.

Each feature has a weighting factor and this weighting factor adjusts adaptively. This makes a feature performs differently in different domains or templates. Experiments show a higher precision of our algorithm than the normal algorithm. However, there were also some shortages. In the future work, we would take the relationship among text fragments into consider. Also, we would research in the entity extraction and go into the information extraction task more deeply.

Acknowledgment. This work is supported by the National Key Technology R&D Program (No.2008BAH24B03) and the National Natural Science Foundation of China (No. 61003254) and Zhejiang Provincial Natural Science Fund of China (No.Y1080130) and the Fundamental Research Funds for the Central Universities.

References

1. Cai, D., Yu, S., Wen, J., Ma, W.-Y.: Extracting content structure for web pages based on visual representation. In: Zhou, X., Zhang, Y., Orlowska, M.E. (eds.) APWeb 2003. LNCS, vol. 2642, pp. 406–417. Springer, Heidelberg (2003)
2. Cai, D., Yu, S., Wen, J.-r., Ma W.-Y.: VIPS: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79 (2003)
3. Carlson, A., et al.: Coupled semi-supervised learning for information extraction. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 101–110. ACM, New York (2010)
4. Chen, L., Ye, S., Li, X.: Template detection for large scale search engines. In: Proceedings of the 2006 ACM Symposium on Applied Computing, pp. 1094–1098. ACM, Dijon (2006)
5. Crescenzi, V., Mecca, G., Merialdo, P.: RoadRunner: automatic data extraction from data-intensive web sites. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, pp. 624–624. ACM, Madison (2002)
6. Junfang, S., Li, L.: Web information extraction based on news domain ontology theory. In: IEEE 2nd Symposium on Web Society SWS (2010)
7. Kushmerick, N.: Wrapper induction: Efficiency and expressiveness. Artificial Intelligence 118(1-2), 15–68 (2000)
8. Labský, M., Svátek, V., Nekvasil, M., Rak, D.: The *ex* project: Web information extraction using extraction ontologies. In: Berendt, B., Mladenič, D., de Gemmis, M., Semeraro, G., Spiliopoulou, M., Stumme, G., Svátek, V., Železný, F. (eds.) Knowledge Discovery Enhanced with Semantic and Social Information. Studies in Computational Intelligence, vol. 220, pp. 71–88. Springer, Heidelberg (2009)
9. Lin, S.-H., Ho, J.-M.: Discovering informative content blocks from Web documents. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 588–593. ACM, Edmonton (2002)
10. Liu, B., Grossman, R., Zhai, Y.: Mining data records in Web pages. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 601–606. ACM, Washington, D.C (2002)
11. Ma, L., et al.: Extracting unstructured data from template generated web documents. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 512–515. ACM, New Orleans (2003)
12. Miao, G., et al.: Extracting data records from the web using tag path clustering. In: Proceedings of the 18th International Conference on World Wide Web, pp. 981–990. ACM, Madrid (2009)

13. Michal Mared, P.P., Spousta, M.: Web Page Cleaning with Conditional Random Fields. Calriers du Central 4, 155–162 (2007)
14. Muslea, I., Minton, S., Knoblock, C.A.: Hierarchical Wrapper Induction for Semistructured Information Sources. Autonomous Agents and Multi-Agent Systems 4(1), 93–114 (2001)
15. Pasternack, J., Roth, D.: Extracting article text from the web with maximum subsequence segmentation. In: Proceedings of the 18th International Conference on World Wide Web, pp. 971–980. ACM, Madrid (2009)
16. Shoubiao, T., Jin, F., Yuan, J.: Web Data Extraction Based on Label Library. In: 2009 WRI World Congress on Computer Science and Information Engineering, (2009)
17. Shuyi, Z., et al.: Joint optimization of wrapper generation and template detection. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, San Jose (2007)
18. Simon, K., Lausen, G.: ViPER: augmenting automatic information extraction with visual perceptions. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, pp. 381–388. ACM, Bremen (2005)
19. Tak-Lam, W., Wai, L.: Adapting Web information extraction knowledge via mining site-invariant and site-dependent features. ACM Trans. Internet Technol. 7(1), 6 (2007)
20. Vadrevu, S., Gelgi, F., Davulcu, H.: Information Extraction from Web Pages Using Presentation Regularities and Domain Knowledge. World Wide Web 10(2), 157–179 (2007)
21. Wang, J., et al.: Can we learn a template-independent wrapper for news article extraction from a single training site? In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1345–1354. ACM, Paris (2009)
22. Wang, J., et al.: News article extraction with template-independent wrapper. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1085–1086. ACM, Madrid (2009)
23. Wong, T.-L., Lam, W.: An unsupervised method for joint information extraction and feature mining across different Web sites. Data & Knowledge Engineering 68(1), 107–125 (2009)
24. Wong, T.-L., Lam, W., Chen, B.: Mining employment market via text block detection and adaptive cross-domain information extraction. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 283–290. ACM, Boston (2009)
25. Xiao, J.-P., Zhang, L.-S., Ren, X.: Web information extraction based on Transductive Support Vector Machine. Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications) 45, 147–149 (2009)
26. Zhai, Y., Liu, B.: Extracting Web Data Using Instance-Based Learning. World Wide Web 10(2), 113–132 (2007)
27. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: Proceedings of the 14th International Conference on World Wide Web, pp. 76–85. ACM, Chiba (2005)
28. Zhao, H., Meng, W., Yu, C.: Mining templates from search result records of search engines. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 884–893. ACM, San Jose (2007)

Infectious Communities Forging
Using Information Diffusion Model in Social Network Mining

Tianran Hu and Xuechen Feng

Hong Kong University of Science and Technology
Department of Computer Science and Engineering
Hong Kong S.A.R
{hutianran001,edward.fxc}@gmail.com

Abstract. This article proposes a new model for clustering individual nodes based on node's interrelation with a real-life mining application. The model is capable of detecting a network topology based on information flow and therefore could be easily extended and applied in a variety of today's research fields. E.g. discover audience group sharing similar attitude, or retrieve authors' academic referencing group or plot active friend society in social networks. An effective algorithm: Boundary Growth Algorithm is proposed through which people can find the underlying structure of networks. Extensive experimental evaluations demonstrate the effectiveness of our approach.

Keywords: Information Diffusion, Social Networks, Community Mining.

1 Introduction

A social group is set of any number of people who share common goals and/or beliefs. Characteristics shared by members of a group may include interests, values, representations, ethnic or social background, and kinship ties. It is believed that the group's behavior is, under certain circumstances, determined by the shared characteristics: in academic field, researchers shared similar academic interests tend to quote each other's article; in social network, people possessing similar hobby would normally form a group. With today's booming of online social network phenomenon, the research on the social group is becoming increasingly popular [1,2,3].

There are two major problems in this field when trying to create a network topology:

Hierarchy problem, or the max boundary problem, that is how to plot the sub-network topology inside an existing group. [7, 4] E.g. certain researchers are more "close" to each other and shared more research interests among a relative large portion of researchers that has the same Artificial Intelligence research interests.

Overlapping problem: how to avoid generate groups of node that is a sub-network of an existing group and contains no desired information itself. [5] E.g. intersection of two social groups from a virtual community does not necessarily provide useful information.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 259–271, 2011.

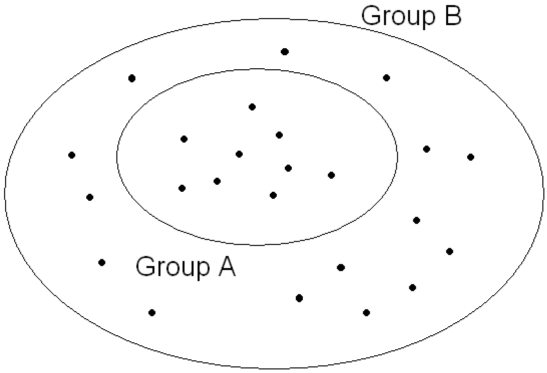


Fig. 1. Hierarchical Structure in Social Group

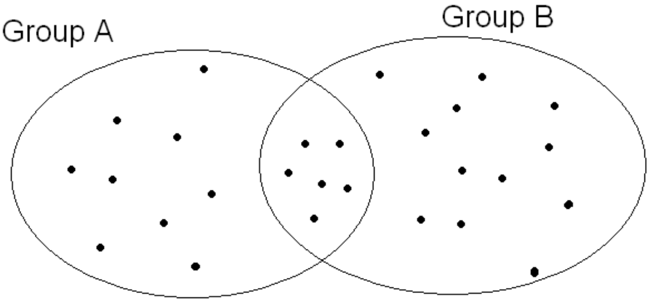


Fig. 2. Overlapping Structure in Social Group

Aside from the topology, information diffusion within the network is yet another intriguing field. [8,9,10] While quite a few progress has been made, there is still a gigantic space for further exploration. Several principle problems such as the interrelation of certain piece of information's boundary when being spread within the network and the network's structure remain unsolved.

In the article, we believe that in addition of the relation with individual node, there is relationship between a piece of information's boundary when being spread within the network and the network's sub-group topology structure. The hypothesis is based on the phenomenon that several nodes in network, especially social network, form a group and it is highly likely that the information would only been spread within the group. In other words, it's been limited by a boundary set by the group. According to the hypothesis, we consider the whole network as a whole, relative loosely organized, group which explain the fact there is certain message that could spread across the entire network.

The proposed method plots the network topology based on the behavior information spread among the network.

One of the problems encountered during the work is the individual node's behavior in social network having certain degree of random aspects, therefore both the input data and the output prediction, or network topology, contain a considerable bias. The problem of inaccurate information delivery could be categorized into two types of situation:

1. For a group, the outsider node may receive the information through the individual with the group. E.g. an author could have his paper being cited by researchers in other research field.
2. For a particular piece of information, the spread may not cover the entire group. E.g. an author does not necessarily cite every other researcher in his research field.

Probability is used in proposed approach to bypass the inaccurate message deliver problem in the network. We consider each message spread as a random event and concentrate only the boundary of the event instead of the source and path of the message. We say if the probability of information to be spread outside to a larger group is small, then the group is considered as a "community". Information is bounded within the group and the member nodes share certain communal aspects.

Based on the assumption, we propose the Boundary Growth Algorithm. The algorithm start from individual node and increment the group side based on information spreading. The algorithm takes the situation of group hierarchy and overlapping into account and could plot a full network topology.

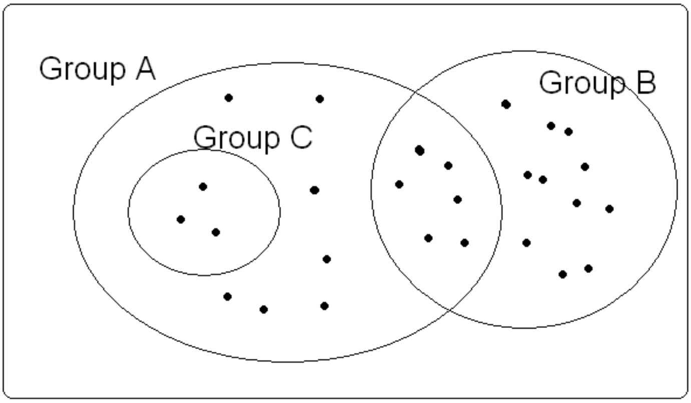


Fig. 3. Ideal Generated Network Groupings

2 Related Work

The proposed algorithm aims to infer community structure in diffusion network. Related work includes diffusion network and community detection.

2.1 Diffusion Network

We consider specifically how new behaviors, practices, opinions, conventions, and technologies spread from person to person through a social network, as people influence their friends to adopt new ideas. The process's mechanism is built on a well known empirical work in sociology: diffusion of innovations [115, 8]. Two of the most influential early pieces in research area to capture such informational effects were Ryan and Gross's study in adoption of hybrid seed corn among farmers in Iowa [9] and Coleman, Katz, and Menzel's study in the adoption of tetracycline by physicians in the United States [10]. In Ryan and Gross's study, farmers were interviewed to determine how and when they decided to begin using hybrid seed corn; the result shows the neighbors' experience of choosing hybrid seed served as the key factor in the decision. In [11], the principle of "homophily" can sometimes act as a "barrier" to diffusion: since people tend to interact with others who share the same social characteristics, while new innovations tend to arrive from "outside" the system, it can be difficult for these innovations to make their way into a tightly-knit social community. This idea shares mutual aspect with our assumption. However in [11], the barrier is used to study diffusion in network itself not the underlying structure of community, which also attracts a lot of attentions recently. J. Leskovec etc. study outbreak detection problem which in [12] is modeled as selecting nodes in a network, in order to detect the spreading of a virus or information as quickly as possible. In [13], Manel and J. Leskovec etc. developed a method for tracking paths of diffusion and influence through networks and inferring the networks over which contagions propagate. Given the times when nodes adopt pieces of information or become infected, identifying the optimal network that best explains the observed infections times. And a linear influence model for predicting which node will influence which other node in the network was proposed by J. Yang etc. [14].

All previous works, though some are similar to our work in this paper, focus on diffusion itself while ignoring the community structure underlying which affect the propagation of information in network most.

2.2 Community Structure Detection

Online social networks are recognized as complex networks which are characterized by high clustering coefficient and short average distance [6]. Data clustering is one of the earliest techniques for community detection, which can be divided into partition-based method such as k-means clustering [15], model-based methods [16], spectral clustering algorithm [19, 17] and hierarchical clustering [47]. Grivan and Newman then introduced several community detection algorithms based on "Centrality" [1]. Modularity-based techniques was introduced in [18] as a measure to evaluate the quality of a set of extracted communities in a network and has become one of the most popular quality functions used for community detection.

However, the majority of works in community detection attempt to discover non-overlapping communities. In [5], X. Wang etc. proposed a method based on tags in social media to find overlapping communities. Ahn et al. [7] apply typical hierarchical clustering to line graph in order to find hierarchical overlapping communities.

The proposed method would be the first time diffusion network concept is extended in community detection field.

3 Problem Formulation

The proposed algorithm aims to overcome the uncertainty from individual behavior by using the information diffusion model. The model is able to focus on information flow behavior within the network therefore bypass the uncertainty brought by individual as well as extend the data space for network structure plotting.

Mathematically, assume we have unit set $U = \{u_1, u_2, \dots, u_n\}$ where u_i (1<=i<=n) is a client that evolves in multiple information diffusion event. We also have information diffusion event $D = \{d_1, d_2, \dots, d_n\}$ where d_i ($1 \leq \mathrm{i} \leq \mathrm{n}$) is the unit the particular diffusion event's "infects", thus $d_i$is a subset of U. also we have community $C = \{c_1, c_2, \dots, c_n\}$ where c_i ($1 \leq \mathrm{i} \leq$n) is a group, or a set of individual unit. So c_i is a subset of U.

In hierarchical and overlapping situations, we may have:

(1) $c_i \subseteq c_j$ ($1 \leq \mathrm{i, j} \leq \mathrm{n}$) for hierarchical sturcture

(2) $c_i \nsubseteq c_j \wedge c_i \cap c_j \neq \phi$ ($1 \leq \mathrm{i, j} \leq \mathrm{n}$) for overlapping structure

Our goal is: given individuals set U and diffusion set D, the proposed algorithm could calculate community set C with hierarchical and overlapping topology structure support.

4 Methodology

4.1 Infection Accounting

Assume v is a set of individuals, thus $v \subseteq U$. If a message M is spread across the entire set, that is, every member in the set get the message, as say this message infects whole set X. Infection accounting $h(v)$ is the total number of messages that infect v. We have:

$$h(v) = \sum_{d_i \in D} f(v)$$

Where for each $d_i \in D$

$$f(v) = \begin{cases} 1 \ (v \subseteq d_i) \\ 0 \ (v \nsubseteq d_i) \end{cases}$$

$h(v)$ represents the active level of the group: the higher $h(v)$, the more information v holds.

Theorem 1. Assume $\mathrm{v}' \subseteq \mathrm{U}$ and $\mathrm{v} \subseteq \mathrm{U}$, if $\subseteq v'$, we have

$$h(v) \geq h(v')$$

Proof:

Since v is a subset of v', for every message M counted in $h(v')$, M infects v. Therefore M infects every member in v. M infects v, it is also counted in $h(v)$. So $h(v) \geq h(v')$.

4.2 Diffusion Proportion

Assume $v' \subseteq U$ and $v \subseteq U$, if $v \subseteq v'$ then diffusion proportion $g(v, v')$ is the probability for a message M spread to $\mathrm{v}'$from v. In other words $g(v, v')$ is the probability for v' to be a subset of d_i where $v \subseteq d_i$.

We have:

$$g(v, v') = P(v'|v) = \frac{h(v')}{h(v)}$$

If for every super set v' of v , $g(v, v')$ is relatively small, obviously, the information in set v is not likely to spread beyond or “bounded” by v.

Definition 1. (Bounded Diffusion Pattern)
Assume $v \subseteq U$ and $V' = \{v'_1, v'_2, \dots, v'_m\}$ and $v \subseteq v'_i$ $(1 \leq i \leq m)$, v is a Bounded Diffusion Pattern if and only if:

$$\forall v'_i (v'_i \in V') \rightarrow g(v, v'_i) < \varepsilon$$

Note that a Bounded Diffusion Pattern is a community in the network.

4.3 Boundary Growth Algorithm

However, for detecting whether a set v is a Bounded Diffusion Pattern, all its supersets needs to be calculated which is computationally infeasible. a simpler and faster model is required.

Theorem 2. Assume $v \subseteq U$, $|v| = a$. if for every superset v' of v where $|v'| = a + 1$, we have:

$$g(v, v') < \varepsilon$$

v is a Bounded Diffusion Pattern.
Proof:
Assume there is an v’s superset v'', $|v''| > a + 1$. Then there is at least one v’s superset v' where, $|v'| = a + 1$ and $v \sqsubset v' \sqsubset v''$. According to Theorem 1, we have $h(v') \geq h(v'')$, thus $g(v, v'') < g(v, v')$. Since $g(v, v') < \varepsilon$, for any superset v'' of X, we have $g(v, v'') < \varepsilon$. So X is a Bounded Diffusion Pattern.

Theorem 2 tells for a set v, we only need to calculate all the superset with cardinality greater than v by one unit to determine if the set is a Bounded Diffusion Pattern.

The Boundary Growth Algorithm is based on theorem 2. It starts at each node of v and build all Bounded Diffusion Patterns includes v. The pseudo code is as follow:

```
Boundary Growth Algorithm
Input: Individuals set U, Diffusion set D
Output: Community set C
Process:
      Set C= $\phi$;
      for each subset v of U;
          Calculate h(v);
      end for;
      for (i = 1; i $\leq$ largest size of diffusion; i ++)
          for  each set v which |v|=i
             boolean flag = ture;
             for each superset v' which |v^' |=i+ 1
   //to check whether v is a bounded pattern, we  only
   //calculate the supersets with cardinality  greater
   //than X by one unit
                  if (g(v,v^' )>$\varepsilon$)
                      flage = false;
                      break;
```

```
        if (flage)
            put  v into C;
           end for;
    end for;
end for;
```

5 Evaluation

The data set Arxiv HEP-TH[1] is used to evaluate the model. Arxiv HEP-TH (high energy physics theory) citation graph is from the e-print arXiv and covers all the citations within a dataset of 27,770 papers with 352,807 edges. If a paper *i* cite paper *j*, the graph contains a directed edge from *i* to *j*. If a paper cites, or is cited by, a paper outside the dataset, the graph does not contain any information about this. The data covers papers in the period from January 1993 to April 2003 (124 months). It begins within a few months of the inception of the arXiv, and thus represents essentially the complete history of its HEP-TH section. Arxiv HEP-TH was originally released as a part of 2003 KDD Cup.

5.1 Setup

We define:

- A link between two authors, if an author cites another researcher's paper in his article.
- An information diffusion event, for an author, each of his paper being cited by any other papers counts as one information diffusion event.

The link distribution from data pool shows a significant heavy-tailed behavior: rather seldom author has a vast citation of his article while the majority of researchers have less than twenty citations, which is about 0.1% of all researchers in the data set. Based on the observation we can conclude while well-known authors and their articles behave as a trivial and dominance factor in the society. Similar to the flooded information entire social network demonstrate the network, as a whole, form a loose community, the obvious pervasive citations of famous articles merely reflects the loosely bound in the data set. This evaluation focuses on the subtle yet particular structures in the network, thus only community structure within researchers of less than fifty citations is considered. After setup, the raw network is trimmed to 27,770 papers with 12,390 authors. Two major parameters are used in evaluation: $h(v)$ threshold and ε value. Different $h(v)$ function threshold is set so that single citation is eliminated to utilize the probability sampling, and ε is used for tune the "sharpness" of the community edge

5.2 Result

5.2.1 Run Time

All set is able to provide a satisfactory run time. Different $h(v)$ threshold and ε value result a monotonously distribution: run time increases along with ε and decreases with $h(v)$ almost linearly.

[1] The data set can be downloaded from url:
`http://snap.stanford.edu/data/cit-HepTh.html`

Table of run time distribution (in ms):

	h(v) = 7	h(v) = 8	h(v) = 9	h(v) = 10
$\varepsilon = 0.3$	278,813	76,781	42,031	26,750
$\varepsilon = 0.5$	292,719	79,875	49,000	29,281
$\varepsilon = 0.7$	310,875	81,516	54,235	30,344
$\varepsilon = 0.9$	323,594	86,046	61,781	32,671

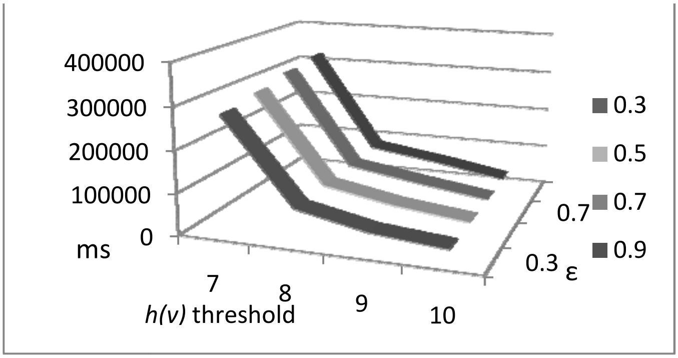


Fig. 4. Performance Distribution

5.2.2 Community Grouping Result

The result demonstrates the difference in grouping caused by Different $h(v)$ threshold and ε value.

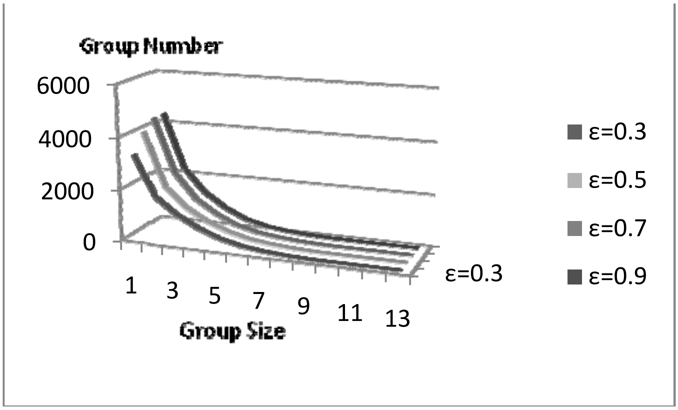


Fig. 5. Difference of group size with different **ε** when $\mathbf{h(v)}$ threshold = 7

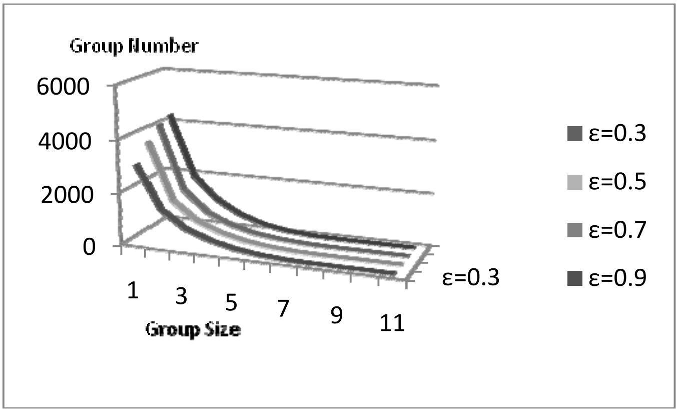


Fig. 6. Difference of group size with different **ε** when **h**(**v**) threshold = 8

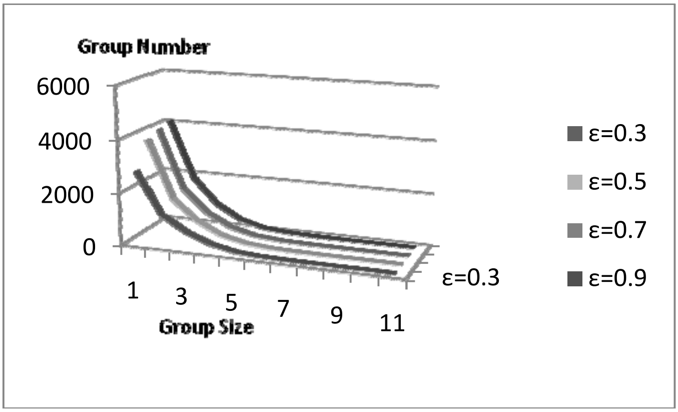


Fig. 7. Difference of group size with different **ε** when **h**(**v**) threshold = 9

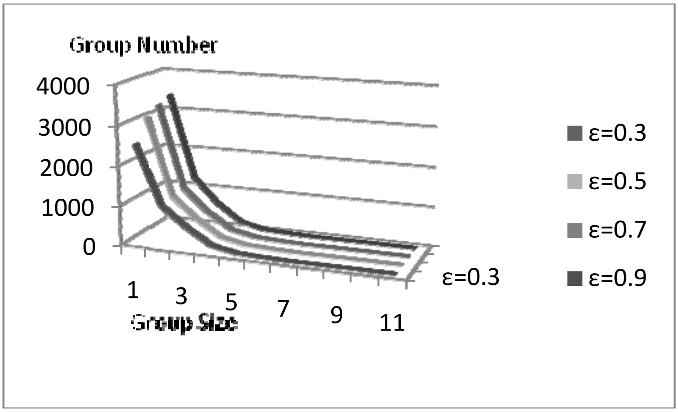


Fig. 8. Difference of group size with different **ε** when **h**(**v**) threshold = 10

Overall, the data shows there is a rather significant amount of small communities with node less than four in the network. However, there is no community with size greater than thirteen in all threshold ε combinations.

As previously proved, the result shows that ε determines whether the existing group is growing, or the "sharpness" of the community edge. Larger ε means a blurred edge, or more combinations nodes. As in the table above, when ε is raised from a range [0.3, 0.9], the valid communities of each size keeping increasing. However, although the total group numbers vary due to the ε value, as the group size grows, the total valid groups tend to become smaller and converge to same result despite the ε value.

The result shows that more than ninety-five percent of the group is of size smaller than five regardless of the choosing of threshold and ε value. Providing the total citations of each individual author are less than twenty, about thirty percent of the total information diffusion event each node participate in is connected and represents a community. Giving the total number of authors is greater than ten thousands; the group size in this network is less than 0.1%. It is confident to say the communities in the network have a rather local and tight relation; or the authors from this journal tend to form a small community and it is not likely that different communities have much communications with respect to article citation.

As result demonstrates, $h(v)$ threshold serves as a filter to eliminate undesired random effects. With higher threshold, more events lesser information diffusion event support is filtered out, thus results in a more concrete topology.

Also, hierarchical community structure is observed during evaluation: when $h(v)$ threshold is set to ten, no community of size ten is generated while one community of size eleven is plotted. This demonstrates that the model is capable of generate hierarchical structure.

6 Conclusion

Due to the extensive involvement of human behavior, the biased input and the unpredictable group behavior are two of the major limitations of today's researches in social network. The existing models for social network structure plotting suffer from poor input quality and performance is limited, particularly in the expression of the hierarchy and overlapping situation among subgroups in network.

The article introduces, illustrate and evaluate a new algorithm that uses information diffusion model to plot a community structure within a network. The model is able to distinguish the hierarchical and overlapping structure among the group and is able to generate the group topology without using any node information or detailed message information.

Due to the little information requirement, the method is suitable to a variety of field to generate group structure. It is believed that this algorithm would contribute in future research and applications in social network mining and beyond.

References

1. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci., 8271–8276 (2002)
2. Hopcroft, J., Khan, O., Kulis, B., Selman, B.: Natural communities in large linked networks. In: Proc. KDD 2003, pp. 541–546 (2003)
3. Newman, M., Barabasi, A.-L., Watts, D.J. (eds.): The Structure and Dynamics of Networks. Princeton University Press, Princeton (2006)
4. Newman, M.E.J.: Detecting community structure in networks. Proc. Natl. Acad. Sci. 99, 7821–7826 (2002)
5. Wang, X., Tang, L., Gao, H., Liu, H.: Discovering Overlapping Groups in Social Media. In: IEEE International Conference on Data Mining, ICDM 2010, pp. 569–578 (2010)
6. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small world' networks. Nature 393(6684), 400–422 (1988)
7. Ahn, Y.-Y., Bagrow, J.P., Lehmann, S.: Linke communities reveal multi-scale complexity in networks (2009)
8. Rogers, E.: Diffusion of Innovations, 4th edn. Free Press, New York (1995)
9. Ryan, B., Gross, N.C.: The diffusion of hybrid seed corn in two Iowa communities. Rural Sociology, 15–24 (1943)
10. Coleman, J., Menzel, H., Katz, E.: Medical Innovations: A Diffusion Study. Bobbs Merrill (1966)
11. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, Cambridge (2010)
12. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriensen, J., Glance, N.: Cost-effective Outbreak Detection in Networks. In: Proc. KDD 2007, pp. 420–429 (2007)
13. Rodriguez, M.G., Leskovec, J., Krause, A.: Inferring Networks of Diffusion and Influence. In: Proc. KDD 2010, pp. 1019–1028 (2010)
14. Yang, J., Leskovec, J.: Modeling Information Diffusion In Implicit Networks. In: 2010 IEEE International Conference on Data Mining ICDM, pp. 599–608 (2010)
15. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (1967)
16. Dempster, A.P., Laird, N.M., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society Series B 39(1), 1–38 (1977)
17. Weiss, Y.: Segmentation using eigenvectors: A unifying view. In: Proceedings of International Conference on Computer Vision (1999)
18. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69, 26113 (2004)
19. Alpert, C., Kahng, A., Yao, S.: Spectral partitioning: The more eigenvectors, the better. Discrete Applied Math. 90, 3–26 (1999)

Appendix

In this section we show the group size of several $h(v)$ threasholds with different ε value obtained from experiments.

Table of group size when $h(v) = 7$ changing ε value:

Group Size	h(v) = 7, $\varepsilon = 0.3$	h(v) = 7, $\varepsilon = 0.5$	h(v) = 7, $\varepsilon = 0.7$	h(v) = 7, $\varepsilon = 0.9$
1	3300	3983	4386	4386
2	1754	1918	2249	2249
3	1148	1177	1313	1350
4	719	725	785	805
5	411	416	421	427
6	211	211	215	219
7	108	108	111	111
8	46	46	46	46
9	15	15	17	17
10	4	4	4	4
11	3	3	4	4
12	2	2	2	2
13	2	2	2	2

Table of group size when $h(v) = 8$ changing ε value:

Group Size	h(v) = 8, $\varepsilon = 0.3$	h(v) = 8, $\varepsilon = 0.5$	h(v) = 8, $\varepsilon = 0.7$	h(v) = 8, $\varepsilon = 0.9$
1	3006	3653	4109	4293
2	1381	1467	1648	1959
3	752	773	835	1067
4	421	424	439	570
5	204	204	208	269
6	89	89	93	109
7	28	28	28	36
8	7	7	7	12
9	5	5	6	7
10	4	4	4	4
11	2	2	2	2

Table of group size when $h(v) = 9$ changing ε value:

Group Size	h(v) = 9, $\varepsilon = 0.3$	h(v) = 9, $\varepsilon = 0.5$	h(v) = 9, $\varepsilon = 0.7$	h(v) = 9, $\varepsilon = 0.9$
1	2782	3759	3959	4101
2	1200	1559	1679	1867
3	641	788	824	897
4	283	332	342	371
5	98	118	118	124
6	30	38	40	45
7	14	16	16	17
8	4	5	5	5
9	2	2	2	3
10	1	1	1	1
11	1	1	1	1

Table of group size when $h(v) = 10$ changing ε value:

Group Size	h(v) = 10, $\varepsilon = 0.3$	h(v) = 10, $\varepsilon = 0.5$	h(v) = 10, $\varepsilon = 0.7$	h(v) = 10, $\varepsilon = 0.9$
1	2522	3053	3255	3399
2	1029	1096	1141	1271
3	561	567	581	625
4	171	173	176	192
5	50	50	50	51
6	19	19	19	22
7	2	2	2	3
8	2	2	2	3
9	3	3	3	3
10	0	0	0	0
11	1	1	1	1

Extracting Dimensions for OLAP on Multidimensional Text Databases

Chao Zhang, Xinjun Wang, and Zhaohui Peng

School of Computer Science and Technology, Shandong University. Jinan, China
kuiste@163.com, {wxj,pzh}@sdu.edu.cn

Abstract. With the amount of textual information massively growing in various kinds of business systems and Internet, there are increasingly demands for analyzing both structured data and unstructured text data. Online Analysis Processing (OLAP) is effective for analyzing and mining structured data. However, while handling with unstructured data, it is powerless. After working on several information integration and data analysis applications, we have realized the defect of OLAP on text data analysis and use technical ways to handle this issue. In this paper, we propose a semi-supervised algorithm to extract dimensions and their members from textual information for the purpose of analyzing a huge set of textual data. We use straightforward measures to express analysis results. Experiment result shows that the extracting algorithm is valid and our approach has a high scalability and flexibility.

Keyword: OLAP, unstructured data, extracting algorithm.

1 Introduction

Since many of business intelligence systems have been utilizing structured data for analysis and decision-making, data warehouses are widely used for organizing and analyzing large amounts of data. A useful technology to exploit data warehouse is the Online Analytical Processing (OLAP) technology [1][2][3], which provides a more flexible representation of multidimensional data in different granularities. The OLAP systems categorize data either as facts with associated numerical measures or as dimensions that characterize the facts. For example, in a recruitment system, recruitment information would be fact, recruitment number would be a measure, work place and recruitment time would be dimensions. OLAP aggregates measures over a range of dimensional members to provide results such as recruitment amount per month in order to analysis changing trend.

On the other hand, with the unstructured text data grows explosively in business systems and Internet, it becomes more and more desirable to extend the traditional OLAP on structured data analyzing as well as text data mining and knowledge discovery [4][5][6]. Generally, these text data exists either in database as data records or in a separate place with links to the documents as description. Here we use multidimensional text databases to represent the form. In order to fully take advantage of all the knowledge, both structured data and unstructured data should be analyzed

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 272–281, 2011.

simultaneously, thus bringing more challenges in knowledge acquisition. Unfortunately, traditional OLAP is powerful in dealing with structured data but faces challenges for analyzing unstructured text data.

In the past, when we want to analyze unstructured text data in a multidimensional data model, we would first choose different dimension member and fetch corresponding documents, and then we use text mining tools to mine these documents in order to acquire knowledge we are interested in. This method has obvious shortcomings, as a decision maker, we must have the professional knowledge of text mining, and, the two steps are taken separately, not in an integration system.

In summary, a more powerful OLAP should integrate text mining with traditional OLAP, and allow the decision makers to drill-down/roll-up/slice/pivot on text dimensions. For example, it would be very useful if we post a query (time=”2010”, jobName=”computer software”) and obtain the recruitment requirements in the domain of computer software, then we drill down into the lower-level categories in jobName dimension, like “software engineer”, to obtain the quality a “software engineer” needed, even we can post two queries (time=”2010”, jobName=” computer software”, education background=”bachelor”) and (time=”2010”, jobName=” computer software”, education background=”master”) to compare the recruitment requirements on different dimension members.

In this paper, we propose a new model to integrate textual information with multidimensional model to achieve the goal of analyzing a huge set of textual documents. This paper describes:

- How we extract dimensions and their members from text data;
- How we efficiently express measures of our analysis model;

2 Related Work

According to the ways on construction of text data hierarchy, we could group the researches of integrating text with OLAP into three categories:

Term as Hierarchy: This kind of approach takes term hierarchy as text dimension which is built to specify term’s semantic levels [7]. Each node in the hierarchy is called a generalized term, represented by a subset of terms. For example, the computer is divided into internal and peripherals, while internal can be divided into memory/CPU/motherboard and so on. When we execute the pull-up operation, given term level L and generalized term v ∈L, add v’s parent node u in L and delete u’s descent nodes from L. The result is a higher term level L’. With the term hierarchy pull-up or push down, the document do the corresponding aggregate operation.

Document Classification as Hierarchy: Most representative of this method are BIW[8] and Polyanalyst[9]. They use classification methods to classify documents into categories and attach documents with class labels. Through the category labels, the analysts would drill-down and roll-up along the text dimension.

Topic as Hierarchy: This method mines the topics of the text data then builds a hierarchy for the topics [10]. In DBPubs[11], the Content Management Systems metadata provide traditional OLAP static dimensions that are combined with dynamic dimensions discovered from the analyzed keyword search result, as well as measures for document scores based on the link between structured data and the documents. For example, a user may post a query:"+venue: SIGMOD +year: 2005", then the system will return the paper topics in SIGMOD 2005, the numbers of each topic are measures. In TopicCube [12], the author use PLSA to mine topics of anomaly event in aviation reports, and then map these topics to a hierarchical topic tree which is defined by domain expert.

The analysis based on topic integrates tightly with OLAP, but the drawback is that, a hierarchical topic tree must be constructed by experts, so this has adverse effects on scalability of systems.

3 System Overview

Figure 1 depicts the system overview. The multidimensional text database stores structured data as well as unstructured text data, the text data is the description of structured data. Take recruitment system for example, the structured data such as time, location, and so on. For obtaining OLAP functionality in analyzing text, our goal is to create a new multidimensional schema by adding additional dimensions into original dimensions. Firstly, structured and unstructured data are stored in multidimensional text databases, they are treated separately, and the structured data are used to construct static dimensions. The second is our core problem, in our approach, text data are processed as dimension extracting and measure computing. We use a semi-supervised learning of dimension-member pairs extracting algorithm to extract dimensions from text, and category dispersion and correlation algorithm to calculate measures. The measures are presented as related phrases frequency which phrases would best reflect job descriptions under different dimension members.

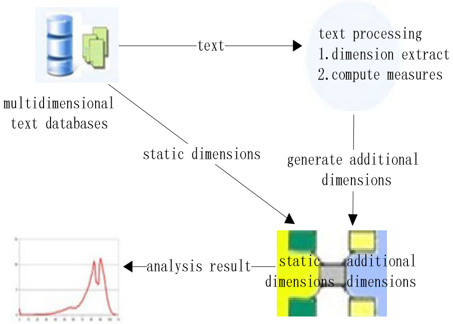
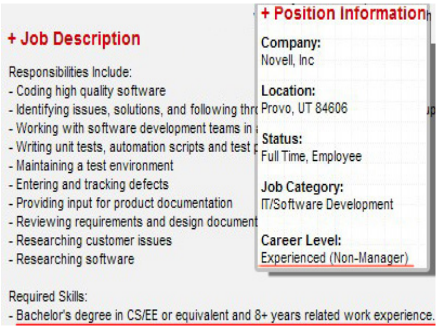


Fig. 1. System overview


+ Job Description

Responsibilities Include:
- Coding high quality software
- Identifying issues, solutions, and following thr
- Working with software development teams in
- Writing unit tests, automation scripts and test p
- Maintaining a test environment
- Entering and tracking defects
- Providing input for product documentation
- Reviewing requirements and design document
- Researching customer issues
- Researching software

Required Skills:
- Bachelor's degree in CS/EE or equivalent and 8+ years related work experience.

+ Position Information

Company:
Novell, Inc

Location:
Provo, UT 84606

Status:
Full Time, Employee

Job Category:
IT/Software Development

Career Level:
Experienced (Non-Manager)

Fig. 2. Example of a recruitment information

4 Extraction of Dimension and Member

The basic idea of our model is to use a semi-supervised learning of dimension-member pairs extraction algorithm to construct Additional Dimension. In other words, the additional dimensions were extracted from document and could be treated as standard dimensions.

4.1 Concept of Additional Dimension

The additional dimensions are constructed based on multidimensional text database. The purpose to do so is extracting dimensions from unstructured documents. Figure 2 shows a piece of recruitment information. Due job description is no theme document, analyze them by mining topic, is senseless.

Definition 1. Additional Dimension is a dimension which extracted from text data and can be treated as standard dimension. Drill-down and roll-up along additional dimension make it available for users to view the data from different granularity of measures.

4.2 A Semi-Supervised Extracting Algorithm

In this section, we describe an approach to extract dimension-member pairs from text [13], like job descriptions, product recommendations and product descriptions etc. This approach uses a semi-supervised algorithm and treats the extraction problem as classification problem. The algorithm uses labeled seeds as training set and needs little user supervision. Firstly, we extract an initial seeds that serves as training data for the semi-supervised classification algorithms. Then, the extracted dimensions and their members are linked by using heuristic rules.

4.3 Seed Generation

The step is to generate labeled seeds for the learning algorithms to learn from. The seeds can be generated in two ways, with small amounts of known dimensions and members by artificially customization, as well as with an unsupervised, automated generate algorithm that extracts seeds from the unlabeled data. Both of these seed generating strategies are aimed to facilitate scaling to other domains. The first way uses small amounts of labeled training data defined by domain experts. Consider the following sentence:

Bachelor's degree in CS/EE or equivalent and 8+ years related work experience.

We use three seed lists: one for education background, one for work experience and one for language requirements. The second way is an unsupervised way. In this method, we consider all bigrams w_iw_j as candidates for dimension-member pairs, where w_i is a candidate member, and w_j is a candidate dimension (e.g., bachelor degree/8+ experience.), although it is not always the case that the member occurs before its dimension in job description. Suppose word w_j occurs with n unique words $w_{1...n}$ in position i, we rank the words $w_{1...n}$ by their conditional probability $p(w_i|w_j)$, $w_i \in w_{1...n}$, where the word w_i with the highest conditional probability is ranked highest.

The highest ranked word w_i are candidates for members for the candidate dimension w_j. Noteworthy is, frequent words and stop words such as *the*, though occur with many different words, they should be abandoned. So, one dimension generally has more than one member and do not occur with a wide range of words. What we interested in is that, few words account for a high proportion of probability. The idea can be divided into two stages: in the first step, use conditional probability to compute candidate dimension by (1), where c is the threshold and $0 < c < 1$; In the second step, compute the mutual information for all candidate dimension-members pairs. If there are a few words that together have high mutual information with the candidate dimension, then we are likely to regard them as the members of dimension. The mutual information can be computed as (2)(3), where λ is a user-specified parameter, where $0 < \lambda < 1$.

$$c = \sum_{i=1}^{k} p(w_i \mid w_j) \quad (1) \qquad p(w, w_{1...k}) = \sum_{j=1}^{k} p(w, w_j) \quad (2)$$

$$cmi(w_{1...k}; w) = \log \frac{p(w, w_{1...k})}{(\lambda * \sum_{j=1}^{k} p(w_j)) * ((\lambda - 1) \& p(w))} \quad (3)$$

After the two steps, some dimension-member pairs maybe generated along with noisy and inaccuracy pairs, in order to ensure correctness, we correct manually.

4.4 Dimension and Member Extraction

After some initial seeds generated, we use the initial seeds as training data and extract dimensions and members from unlabeled data. The extraction issue can be regarded as a classification problem because of each word or phrase can be classified as dimension or member or neither. This can be viewed as a semantic tagging process. We use four classes to classify words into: dimension, member, unassigned or neither. Initially, each word is labeled unassigned as default, if the unassigned words match the labeled data, then we assign it as the matched label. On the other hand, if the unlabeled words do not match the labeled data, we input them to the classification algorithm.

The algorithm we use is co-EM which is a multi-view semi-supervised learning algorithm [14]. The labeled words are used as training data for co-EM that classifies each word to the unlabeled data as dimension, member, or neither .The co-EM combines features from both co-training and Expectation-Maximization (EM), it is an iterative algorithm just like EM, but uses the feature split present in the data, just like co-training. The features we use for classification are the words of each unlabeled data item, the surrounding n words, and their corresponding parts of article. The data are expressed in two views; view1 includes each word itself, along with part-of-speech assigned by semantic analysis tools, here we use Stanford Log-linear Part-Of-Speech Tagger [15]. View2 is the context size n, that is to say, n/2 words ahead or after the word in view1. For example, view1 includes the word years, n=3, view 2 has CS/EE, equivalent, 8+, related, work, experience while interference words have be removed.

In the initial state, view1 uses the labeled data only. Then the classifier labels all unlabeled data item probabilistically. The view2 is then trained using the original labeled data along with the unlabeled data with the labels provided by the view1 classifier. Similarly, the view2 classifier relabels the data labeled by the view1 classifier, and this process iterates for several times until the classifiers converge.

If a word$_i$ in view2 does not match the labeled training data, use view1 classifier to train view2 is done by (4), otherwise, the initial labeling is used.

$$p(c_k \mid view2_i) \propto p(c_k) * p(view2_i \mid c_k) \quad (4)$$

In this formula, P (c_k) is the class probability of class ck which is estimated using the current labels in another view. View2$_i$ is the word$_i$ in view2, P (view2$_i$|c_k) is the word probabilities which estimated using the current labels in another view as well as the concurrence times the view2 data item view2$_i$ and view1 data item occurs. The opposite direction is done similarly. The final probability distributions <view1$_i$, view2$_j$> can be assigned as (5):

$$p(c_k \mid <view1_i, view2_j>) = \frac{p(c_k \mid view1_i) + p(c_k \mid view2_j)}{2} \tag{5}$$

The tagged word with high probability as dimensions or members will be extracted as the following rules.

4.5 Find Links between Dimensions and Members

The former section assigns a probability distribution over all the labels to each data item, and then we need to find links between dimensions and members to form dimension-member pairs. We use a heuristic method to do this:

Rule1: Link if dimensions and members match a seed pair.

Rule2: If the words of the same label have a threshold exceeding correlation scores, merge them.

Rule3: Link if dimension and member exceed a co-location threshold.

Rule4: Link if dimension and member are adjacent.

Rule5: If the data item labeled as dimension appears frequently or if the unlabeled data item consists of only one word, extract binary dimensions. Like experienced, its member is true or false, but the member not evidently indicated.

4.6 Generate Additional Dimension

After the steps are done, we would get the dimension-member pairs as bachelor degree, master degree, 5 years experience, etc. and classify descriptions according to dimension members. The left work becomes simple: we could construct dimensions such as degree dimension, work experience dimension and so on.

5 Measure Calculation

Measure calculation is also an import work, which portrays the characteristic intuitively. When taking a different dimension of member, the corresponding documents are doing aggregation operations. Common text characteristic evaluation functions are method based on characteristic of word frequency/document frequency, document information extraction, information gain, mutual information gain, etc.

Characteristic of Word Frequency/Document Frequency: statistics the word/document frequency of a characteristic word. According to the threshold decide whether to retain or remove the characteristic. Disadvantages: rare words may contain important information; lack of theoretical basis.

Information Gain: the average information of a category when a document includes the characteristic, defined as the entropy difference when a feature appears in the text before and after. Disadvantages: When the sample distribution and characteristic distribution are not balance, there is data sparseness problem, which greatly reduce the effectiveness of information gain.

Mutual Information Gain: in statistics, it is used to characterize the correlation between two variables. The mutual information gain between characteristic t and document category c is defined as: MI(t,c)=log(p(t|c)/p(t)). Disadvantages: impacted much by the critical characteristic probability, when the p(t|c) of characteristic are equal, the rare words have a higher score than average score, therefore, text feature probability that much difference of mutual information value is not comparable.

In our model, we have explicit classification according to different dimension members. For example, the document belongs to year=2010.04, location=Beijing, education background=bachelor), we consider that the importance of characteristic not only with the times it appears in document, but also related to its category degree of correlation and its distribution in category. We proposed an improved characteristic calculating algorithm based on category degree of correlation and category degree of dispersion.

Category Correlation: The characteristic words should appear in one category or several categories, not scattered appear in various documents.

$$\rho_{ij} = \sum_{k=1 \text{and } k\neq j}^{m} \left[\frac{DF_n(f_i,c_j) - DF_n(f_i,c_k)}{\sum_i DF_n(f_i,c_j)} \right]^2 \tag{6}$$

m is the number of feature words, $DF_n(f,c)$ is the number of documents that characteristic word f appears at least n times in category c. The bigger ρ_{ij}, the higher degree of correlation between feature word f_i and category c_j.

Category Dispersion: Once a characteristic word is important to a category, it would widespread distribute in this category other than frequently appeared in this category of individual text. We use S_{ij} to express it.

$$S_{ij} = \frac{\mathrm{DF}_n(f_i,c_j)}{\sum_{k=1}^{m} DF_n(f_k,c_j)} \times \frac{TF(f_i,c_j)}{\sum_{k=1}^{m} TF(f_k,c_j)} \tag{7}$$

TF is the number a characteristic appears in a document. The smaller S_{ij} indicates characteristic word f only appears in category c individually.

So the comprehensive importance is:

$$W_{ij} = a * \rho_{ij} + b * S_{ij} \tag{8}$$

where a, b are coefficients given by user. In our experiment, both of them are set to 0.5.

By (8) we could compute eigenvalue of each word and select the highest scoring characteristic as our measure.

6 Experiments

In this section, we present evaluation of our OLAP analysis model. The dataset we used was extracted from http://www.51job.com/, which contains recruitment descriptions as unstructured data and structured data like location, job name and so on. The data set contains 77470 records. Two dimensions were extracted from job descriptions: experience and education background.

In extracting algorithm, we automatic generated 82 seeds. Table 1 shows the example of the automatic generated seeds. After seeds generated automatically, we still need to manually correction. For example, we might do not need the dimension java application, then we delete it form candidate seed list.

Table 1. Example of automatic generated seeds

source phrases	dimension	member
java application	application	java
experienced	experienced	true
5+ years experience	experience	5+ years
senior java developer	java developer	senior
master’ s degree	degree	master’s
minimum 3 years client server experience	experience	3 years

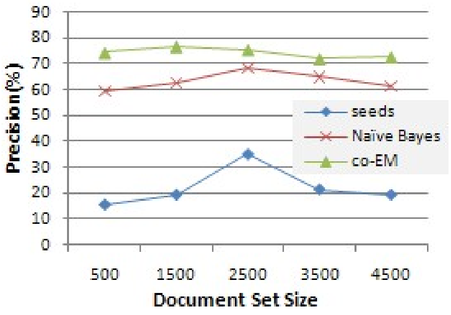


Fig. 3. Algorithm evaluation on precision

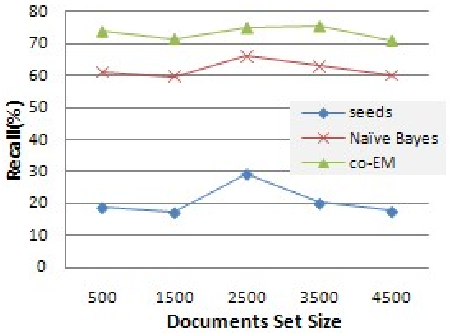


Fig. 4. Evaluation on recall

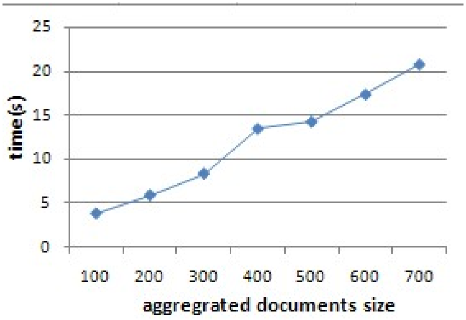


Fig. 5. Spending time during calculating

6.1 Extraction Algorithm Evaluation

Our extraction algorithm has two steps: (1) Generate seed list base on probability. (2) Use co-EM for dimension extraction. As figure 3 and figure 4 shows, recall and precision differs much in different algorithm. But the same algorithm in different document scale recall and precision rate differs less. The seed generation algorithm performs poorly, only few seeds were correctly extracted, the efficiency is very low. But as we know, it is not designed to have a high recall rate, because we have to correct the seeds manually. Also we can know that the co-EM performs better than basic classification algorithm- Naïve Bayes.

6.2 Measure Calculating Time

Measure mining evaluation: in (8), we set a=0.5, b=0.5. The setting is tested by experience and experiment. We do not know that beforehand. Figure 5 shows the spending time during calculating, with the size of the aggregated documents rising, the cost time is almost linear growth.

6.3 Case Study

Table 1 gives a typical example of our analysis model. The Time, Location, Job Name are standard dimensions which have been omitted, Work Experience and Education Background are additional dimensions extracted from document. The measures are most reflecting characteristics of one category. We can get difference between descriptions of bachelor degree and master degree under the same job position as *systems analyst.*

Our analysis model not only expands the dimension for analysis, but also provides reference for analysis.

Table 2. A typical example of our model

Additional Dimension		Measure
Work Experience	Education Background	
+All	-All	SQL, algorithm, logical thinking ability, system optimization, coding
+All	bachelor	Daily reports, business rules, J2EE, SQL, requirement analysis, system analysis
	master	Data mining, calculation method, business statistics, commercial index, Linux
	…	…

7 Conclusion

In this paper, we combine structured data analysis with unstructured data mining. We propose a new analysis model of OLAP which extracts dimensions from text. This expands the range of analysis, provides meaningful measures for analysis. We use a semi-supervised extracting algorithm and a new measure calculating method based on category correlation and category dispersion. We use a materialize strategy which generate candidate views to be materialized and this strategy could reduce the calculation cost of view benefit. Experiment result shows that the extracting algorithm is valid and our approach has a high scalability and flexibility.

Acknowledgement. We sincerely thank the anonymous reviewers for their comprehensive and constructive comments. This research is supported by the National Natural Science Foundation of China under Grant No.61003051, National Key Technologies R&D Program No.2009BAH44B02, Natural Science Foundation of Shandong Province of China under Grant No. ZR2010FM033, Shandong Distinguished Middle-aged and Young Scientist Encouragement and Reward Foundation under Grant No. BS2009DX040.

References

1. Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J.F., Ramakrishnan, R., Sarawagi, S.: On the computation of multidimensional aggregates. In: VLDB, pp. 506–521 (1996)
2. Chaudhuri, S., Dayal, U.: An overview of data warehousing and olap technology. SIGMOD Rec. 26, 65–74 (1997)
3. Gray, J., Bosworth, A., Layman, A., Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. In: ICDE, p. 152 (1996)
4. Wu, T., Xin, D., Mei, Q.: Promotion analysis in multi-dimensional space. In: VLDB 2009 (2009)
5. Inokuchi, A., Takeda, K.: A Method for Online Analytical Processing of Text Data. ACM, New York (2007)
6. Baid, A., Balmin, A., Hwang, H.: DBPubs: multidimensional exploration of database publications. ACM, New York (2008)
7. Lin, C.X., Ding, B., Han, J., Zhu, F., Zhao, B.: Text Cube: Computing IR Measures for Multidimensional Text Database Analysis. In: ICDM (2008)
8. Cody, W.F., Kreulen, J.T., Krishna, V., Spangler, W.S.: The integration of business intelligence and knowledge management. IBM Syst. J. 41, 697–713 (2002)
9. Megaputer's polyanalyst, `http://www.megaputer.com/`
10. Yu, Y., Lin, C.X., Sun, Y.: iNextCube: Information network-enhanced text cube. ACM, New York (2009)
11. Simitsis, A., Baid, A., Sismanis, Y., Reinwald, B.: VLDB 2008 Multidimensional Content eXploration (2008)
12. Zhang, D., Zhai, C., Han, J.: Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases. In: SDM (2009)
13. Liu, Y.: Semi-Supervised Learning of Attribute-Value Pairs from Product Descriptions. In: IJCAI 2007 (2007)
14. Brefeld, U.: Co-EM support vector learning. In: Conference on Machine Learning (2004)
15. Stanford Log-linear Part-Of-Speech Tagger,
`http://nlp.stanford.edu/software/tagger.shtml`

A Conceptual Framework for Efficient Web Crawling in Virtual Integration Contexts*

Inma Hernández, Hassan A. Sleiman, David Ruiz, and Rafael Corchuelo

University of Seville,
Seville, Spain
{inmahernandez,hassansleiman,druiz,corchu}@us.es
http://www.tdg-seville.info

Abstract. Virtual Integration systems require a crawling tool able to navigate and reach relevant pages in the Web in an efficient way. Existing proposals in the crawling area are aware of the efficiency problem, but still most of them need to download pages in order to classify them as relevant or not. In this paper, we present a conceptual framework for designing crawlers supported by a web page classifier that relies solely on URLs to determine page relevance. Such a crawler is able to choose in each step only the URLs that lead to relevant pages, and therefore reduces the number of unnecessary pages downloaded, optimising bandwidth and making it efficient and suitable for virtual integration systems. Our preliminary experiments show that such a classifier is able to distinguish between links leading to different kinds of pages, without previous intervention from the user.

Keywords: Crawlers, Web Navigation, Virtual Integration.

Virtual Integration aims at accessing web information in an automated manner. The virtual integration process starts with queries, in which users express their interests and information needs, and its goal is to obtain information relevant to those queries from the web (probably from different sites), and present it uniformly to the users. By relevant page, we mean a page that contains the information required to answer a user query.

Automated access to the web requires a crawler, that is, a tool able to navigate through web sites, looking for relevant pages, from which to extract the information that is returned to the user. Note that this process is online, which means that bandwidth and efficiency are important issues regarding virtual integration crawling [8], and downloading a minimum number of irrelevant pages is mandatory.

In the design of a virtual integration crawler we consider some requirements: first, the crawler must be able to fill in and submit forms, to access pages in

* Supported by the European Commission (FEDER), the Spanish and the Andalusian R&D&I programmes (grants TIN2007-64119, P07-TIC-2602, P08-TIC-4100, TIN2008-04718-E, TIN2010-21744, TIN2010-09809-E, TIN2010-10811-E, and TIN2010-09988-E).

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 282–291, 2011.

the Deep Web; then, as we said before, the crawler must be efficient, that is, it should minimise bandwidth usage and number of irrelevant pages downloaded. To accomplish this, features for classification must be located outside the page being classified; finally, creating large labelled training sets is burdensome for the user, so we focus instead on training the crawler using an unlabelled set obtained automatically.

Our goal in this paper is to present a conceptual framework that supports the design of crawlers supported by URL-based classifiers. These crawlers determine a page relevance from its URL without having to download it, which reduces the bandwidth and makes them efficient and suitable for virtual integration systems. Even though there are other crawling techniques and tools available that improve traditional crawlers efficiency, our proposal is different, since it is based on link classification to avoid downloading irrelevant pages. We focus on crawling web sites that are designed following a certain navigation pattern, which is the most common pattern in the Web [12]. This pattern consists on a form page that allows issuing queries, followed by a hub page that contains a list of responses to the queries, each of them containing links that finally lead to detail pages.

The rest of the article is structured as follows. Section 2 describes the related work; Section 3 presents the conceptual framework proposed to solve the aforementioned problem; finally, Section 4 lists some of the conclusions drawn from the research and concludes the article.

1 Related Work

Next, we describe the related work in the area of web crawling, enumerating the existing theoretical techniques, and analysing them according to the previous requirements.

1.1 Crawling Techniques

Crawlers are designed considering different requirements, according to their purpose. We distinguish between traditional crawlers, recorders, focused crawlers, and others.

Traditional crawlers [18] collect as many pages as possible from the Web, starting at a given entry point and following links until they meet some stopping conditions. These crawlers have many applications, being the most obvious to create a cache of all visited pages so that a search engine can index them later. Other typical crawling tasks are related to web site maintenance, like validating HTML or performing stress testing.

A recorder is a crawler in which each navigation step is defined by the user. Some examples of recorders are [3], [4], [6], [15], [21]. All of them rely on the user to define which links should be followed in every step, what forms to be filled, and which words to be used for that purpose, so that the crawler reaches exactly the pages targeted by the user. To help the user in the definition tasks, many proposals include a supporting graphical interface [3], [15]. However, in non-GUI proposals, the user has to know HTML code details, and define a step-by-step script, including the references to fields that must be filled in, the actions needed to submit the form (i.e., clicking on a button or a link), and the indication of what links to be followed. In both cases, the user has to provide the values for the different form fields.

Focused crawlers [1], [7], [5], [14], [16], [17] are crawlers which retrieve all pages belonging to a certain topic. They are used, for example, to create a corpora of documents for information retrieval purposes. Their behaviour is similar to that of a traditional crawler, but every retrieved page is analysed to check if it belongs to the topic, usually, with the help of a content-based web page classifier. If the page belongs to the topic, all its links become new seeds for the crawler. Otherwise, the page is judged not relevant and discarded.

Content-based classifiers use either: i) features in the page itself, often a bag of words; or ii) features that are in neighbour pages (pages that either link to or are linked by the target page), like words in the text surrounding the link, the link anchor, and the URL itself. Getting features from the linking page for classification avoids downloading the page beforehand. If the text surrounding the anchor or the anchor text itself contain descriptive words, it is possible to decide the page topic prior to downloading it.

Other crawlers consider different selection criteria for retrieved pages. For example, features like page structure or its location inside the web site directory. Furthermore, they are automated, requiring little intervention from the user, which distinguishes them from recorders. Some examples are [12], [13], [20].

1.2 Analysis

In Table 1, we present a comparison of existing crawling techniques, regarding the following requirements:

- Form filling: either user defined (UD) or applying intelligent techniques. As for the source of keywords for filling, it can be either values from a database (DB), provided by the user (UD), or extracted from the site itself (FA, TDF-IDF) (Column 1)
- Efficiency: Optimisations made to the crawler to reduce the number of irrelevant pages retrieved (Column 2)
- Features: whether they are obtained from the page itself (TP) or from the pages linking to it (LP) (Column 3)
- Training set: labelled (L) or unlabelled (U) (Column 4).

2 Conceptual Framework

We first present an overview of the framework, as shown in Figure 1. Then, we present the details of each module, including a definition of its responsibilities, an example of a typical use case, a list of the possible issues that should be considered in the design, and a discussion of the alternative solutions.

Table 1. Related work (UD = User Defined, ML=Machine Learning, FA=Frequency Analysis, LP=Linking Pages, TP=Target Pages, L=Labelled, U=Unlabelled)

CRAWLING TECHNIQUE	PROPOSAL	FORM FILLING		EFFICIENCY	FE AT	TS
		TECHNIQUE	KEYW			
TRADITION.	Ravaghan01	Matching fields - values	DB	-	-	L
	Barbosa04	Automated	FA	-	-	U
	Madhavan08	Automated	TF-IDF	-	-	U
	Ntoulas05	Query selection algorithm	DB, FA	-	-	U
FOCUSED	Chakrabarti98	-	-	-	LP	U
	Chakrabarti99	-	-	Content classifier & hubs and authorities location	TP	L
	Aggarwal01	-	-	Multiple features classifier	LP	L
	Mukherjea04	-	-	Nearness in directory structure and discarding useless directories	TP	L
	Pant05	-	-	Pre-crawled classifiers	TP	L
	Barbosa05	-	-	Link classifier	LP	L
	Pant06	-	-	Link context classifier	LP	L
	Assis07	-	-	No training, genre & topic classifier	TP	-
	Partalas08	-	-	Reinforcement Learning	TP	L
RECORDER	Anupam00	UD	UD	UD	-	-
	Davulcu99	UD	UD	UD	-	-
	Pan02	UD	UD	UD	-	-
	Baumgartner05	UD	UD	Links matching XPATH given by user	-	L
	Blythe07	UD	UD	Links matching model from users actions	-	L
	Bertoli08	Discard password and keyword forms	-	Links matching model from users actions	-	L
	Wang08	Automated	UD	Paradigm Page-Keyword-Action	-	L
OTHERS	Liddle02	Default Query (empty fields)	-	-	-	
	Lage04	Matching fields - values	UD	Object-rich page classifer; detect Next links	-	
	Vidal07	Automated	FA	Structural classifier	TP	L

A virtual integration process starts with an enquirer, which translates the user interests into queries that are issued to forms. Usually, responses to queries are hub pages, lists of results ordered and indexed, each of them showing a link to another page with detailed information. Relevant pages, when found, are passed on to the information extractor, which obtains and structures the information, that is returned to the user.

We distinguish two phases: the training and the normal execution phase. In the training phase, the keyword manager and form filler focus on obtaining automatically a set of links from the site under analysis, which is later used to train the classifier. In the latter phase, the form filler is used to reach pages behind the forms, and then the crawler uses the trained classifier to select which links to follow. In this paper we focus on the training phase of the framework, namely in the setup and classifier modules.

The only requirement for the training set is to be representative of the site under analysis, hence it is extracted from hub pages, which contain a high number of links in comparison with the rest of pages in any site. Furthermore, they are pages linking directly to pages containing the relevant information, so they assure that we have examples of links leading to relevant pages.

2.1 Keyword Manager

The keyword manager is responsible for finding a list of keywords that allow to obtain a representative collection of links when performing the corresponding searches in a given web site. As an example, to obtain a collection of links from Amazon.com, which offers a variety of products, the keyword manager chooses

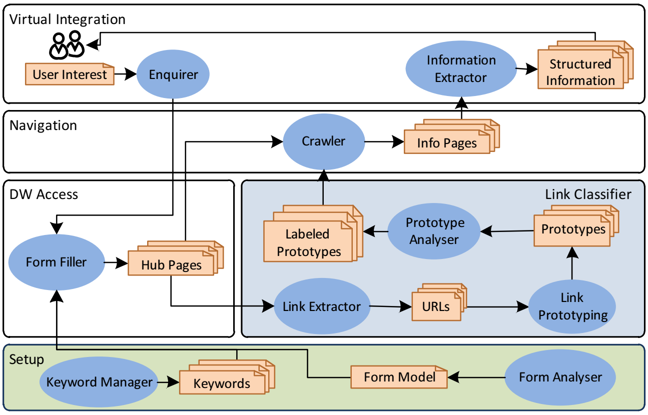


Fig. 1. Conceptual framework diagram

a list of the most common English words [9]. Instead, another site like Microsoft Academic Search belongs to a more specific domain, so a list of the most cited authors, for example, would be more useful for this purpose.

The main concerns in this module are related to the language and type of words that are accepted by each site, specially stop words. For instance, stop words tend to have a higher frequency, and they usually yield a higher number of links, but not every site takes stops words into account. Consider Wordpress.com, which is unable to find any result related to the keywords ’a’ or ’the’, while the same words in Youtube.com yield respectively 32,800,000 and 40,000,000 results. Furthermore, stop words may deviate the search and deteriorate results. The lexical type of word must also be considered, given that verbs are not as frequent as nouns, for example, so they may yield a smaller number of results. Other important factor is the domain to which the site belongs, since it defines a specific vocabulary.

The simplest solution consists of finding a public or well-known corpus, like the Oxford English Dictionary [9]. In some specifical domains, it may be more difficult to find such a list. For example, to search in Apple Store, we need to find a list of the most frequent words in English related to media and technology. If we try to use the list of English most common nouns, we find that for ’week’ (17th position), the Apple store is unable to find any related results. However, in a more general site like Amazon, the same keyword yields 61,494 results.

The last resort is to use the pages of a site to extract the list of keywords, by performing a frequency analysis of the words in the site pages.

2.2 Form Analyser

The form analyser is responsible for visiting a given site, obtaining all the information about the forms and fields that it contains and using that information to build a site form model. For example, to extract information from Amazon, the form analyser opens the Amazon home page, where it finds the following form:

```
<form action="searchAction" name="site-search">
 <select name="url" id="searchDropdownBox" >
  <option value="aps">...</option>
 </select>
 <input type="text" id="twotabsearchtextbox"
name="field-keywords" />
 <input type="image"
src="http://images-amazon.com/images/..."/>
</form>
```

The analyser then obtains a model that includes the form, with a name attribute with value site-search, and no id attribute, its three fields, and its submission method, which consists of clicking over the image.

One of the main issues to be solved by the analyser is the lack of standardisation in forms and fields identification. The analyser (and later, the form filler) has to deal with the problem of referencing that HTML element for later processing. HTML standard includes identification attributes for each element ("id" and "name"), but actually in some web sites we find form elements with none of them. In the latter case, the analyser needs to use a different location strategy, usually in the form of an XPATH expression.

However, although XPATH is flexible and allows to define an expression for every element, this expression can be hard to understand and handle, and also sensitive to small changes in the HTML page. Of course, changes in an element id or name can also invalidate the former location strategy, but it is more usual to change an HTML page by inserting a new paragraph, or deleting an image, than to alter an element id or name attributes.

Youtube.com is a good example of the variety of ways to identify form elements. In its home page we can find, amongst others, a form with id and no name: <form id="masthead-search" action="/results" >, a form with name and no id <form name="logoutForm" action="/">, and a form without name nor id <form action= "/addtoajax">

There are many form modeling proposals, ranging from simple models that just keep a record of all the fields to more complex models that add semantics to each field, analysing field tags [2], [10], [12], [18] surrounding text, identifying mandatory fields [19] or relationships between fields [10].

2.3 Link Extractor

The link extractor is responsible for the extraction of links from hub pages. In the Amazon example, for every page retrieved by the form filler, the extractor analyses all the anchors in the page, including the following:

```
<a id="navLogo" href="/ref=logo">
<a href="http://www.amazon.com/Time-1-year-
auto-renewal/dp/ref=sr1?ie=UTF8"><img
src="http://images-amazon.com/AA160.jpg" /></a>
```

```
3: <a href="http://www.amazon.com/Time-1-year-
   auto-renewal/dp/ref=sr1?ie=UTF8">Time (1-year
   auto-renewal)</a>
4: <a href="http://www.amazon.com/Blind-Side
   -Sandra-Bullock/dp/ref=sr7?ie=UTF8&s=dvd">The Blind Side</a>
5: <a href="http://www.amazon.com/Time-1-year-
   auto-renewal/product-reviews/ref=sr1img?ie=UTF8"/>
6: <a href="javascript:void(0);">Let us know</a>
7: <a href="http://ad.doubleclick.net/clk;..."> <img
   id="nav-site" src="http://images -amazon.com/img2.jpg"></a>
8: <a href="http://www.amazon.com/gp/offer-
   listing/ref=olp?ie=UTF8&condition= new">79 new</a>
9: <a href="http://www.amazon.com/gp/offer-
   listing/ref=olp?ie=UTF8&condition= used">43 used</a>
```

The main issue for the extractor is that a single URL may be written in different formats, either in relative or absolute form. For example, in Amazon, link 1 can be written http://www.amazon.com/ref=logo, /ref=logo or ./ref=logo. Therefore, the system needs to transform all links to their absolute form.

A related issue are links that do not lead to a page, e.g, links to JavaScript functions (in the previous example, line 6), and hence should be discarded, as they are not useful for our purposes. Duplicated links have to be discarded as well.

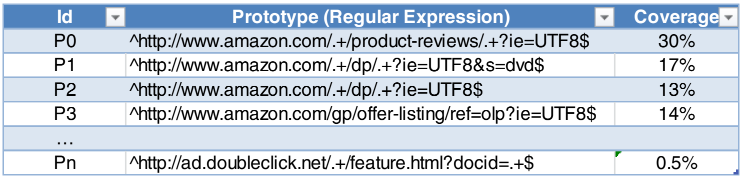
2.4 Link Prototyping

Its goal is to build a collection of prototypes from the set of extracted links. Each prototype represents a subset of URLs, hence it is defined by a regular expression, and the set of all prototypes is later used to build a link classifier. In traditional machine learning approaches, prototype based classifiers classify elements by computing the distance between the element to be classified and each prototype, and assigning the element to the cluster of the prototype whose distance is lower. In our case, instead, whenever a link matches the regular expression of a prototype, it is assigned to its related cluster. In addition, for each prototype, a coverage value is estimated that gives a hint about the importance of the cluster, by counting the number of URLs in the link training set that match its regular expression. In the Amazon example, analysis of training links yields the prototypes in Table 2.

The link analysis is supported by a tokeniser, which parses every URL and splits it into its different components according to RFC 3986. Sometimes URLs include special characters, spaces and other symbols that make it difficult parsing URLs. Furthermore, URL query strings contain parameters, which may be optional or mandatory, and which may be arranged in different orders. The generator has to detect this in order to make a more accurate regular expression.

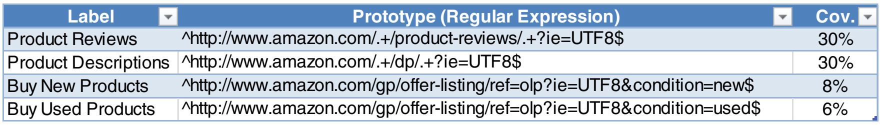
URL rewriting makes things worse by eliminating the structure of a query string in exchange for 'friendly' URLs with a better readability. URL rewriting is not a standard procedure, but only a concept that is being adopted lately by many popular web sites, each of them defining their own friendly URL format.

Table 2. Prototypes obtained from Amazon.com, before user labelling



Id	Prototype (Regular Expression)	Coverage
P0	^http://www.amazon.com/.+/product-reviews/.+?ie=UTF8$	30%
P1	^http://www.amazon.com/.+/dp/.+?ie=UTF8&s=dvd$	17%
P2	^http://www.amazon.com/.+/dp/.+?ie=UTF8$	13%
P3	^http://www.amazon.com/gp/offer-listing/ref=olp?ie=UTF8$	14%
...		
Pn	^http://ad.doubleclick.net/.+/feature.html?docid=.+$	0.5%

Table 3. Labelled prototypes obtained from Amazon.com



Label	Prototype (Regular Expression)	Cov.
Product Reviews	^http://www.amazon.com/.+/product-reviews/.+?ie=UTF8$	30%
Product Descriptions	^http://www.amazon.com/.+/dp/.+?ie=UTF8$	30%
Buy New Products	^http://www.amazon.com/gp/offer-listing/ref=olp?ie=UTF8&condition=new$	8%
Buy Used Products	^http://www.amazon.com/gp/offer-listing/ref=olp?ie=UTF8&condition=used$	6%

2.5 Prototype Analyser

Once the prototyping is complete, the prototypes are analysed and improved so that the result is more accurate. Finally, the analyser helps the user to assign a label to each prototype, defining the semantic concept contained in the links of the cluster that the prototype represents. Moreover, the user selects from the set of prototypes those representing relevant concepts, which will be considered during later crawling.

In the Amazon example, after processing the prototypes the analyser outputs, amongst others, the labelled prototypes included in Table 3.

There are some ways to improve prototypes accuracy, mainly: prototype joining, prototype splitting and prototype discarding.

Joining two different prototypes representing the same concept results in a single prototype with a more general regular expression. For example, prototype P1 is composed of pages with DVD products information, while P2 represent pages about any other type of products. They are joined to form Product Description prototype in Table 3.

Splitting a prototype results in two smaller and more cohesive prototypes. For example, in Amazon prototype P3 includes both links to buy new and used products. The analyser splits this prototype into two different prototypes: one for buying used products and one for buying new products.

URLs that appear only a few times in the training set and whose format is completely different from the other URLs, lead to a prototype with low coverage, which are seldom helpful for the crawler, (e.g., advertising URLs), so they can be discarded without diminishing the classification power of the prototype set. For example, the last prototype in Table 2, whose coverage is low in comparison with the others (0.5 %), is excluded from the labelled prototype set in Table 3.

3 Conclusions

In this paper, we present a conceptual framework for designing crawlers supported by a prototype based link classifier, that classifies pages according to their URL format without downloading them beforehand. Parting from an unlabelled set of links, a set of prototypes is built, each of them representing one of the different concepts embodied in a particular web site. Then, the user selects those concepts that are relevant, and the crawler uses the related prototypes to reach only pages containing those concepts. With respect to the requirements we mentioned in section 1, we observe the following:

Efficiency: Our proposal classifies web pages depending on the link URL format, it is not only efficient, but also generic and applicable in different domains.

Traditional crawlers browse the whole site, retrieving all pages, and spending a significant time and bandwidth while downloading them. Focused crawlers retrieve pages belonging to a topic more efficiently than traditional crawlers do, but still a page has to be classified to know if the crawler must follow that path, and that requires the page to be downloaded in most cases.

In general, using features located on a page to classify it requires downloading it previously, which results in wasted bandwidth. There are, some proposals that classify pages according to the anchor text and text surrounding the link in the referring page. However, not all sites include in their links and their surroundings words useful for classification. The same problem of specificity can be noted in ad-hoc techniques (classifiers designed for a specific site), and also in recorders.

Form Filling: Our goal is to adapt existing form modeling proposals and integrate them into our crawler. As we explained in Section 1, form filling has been studied thoroughly in the literature, so it is not our matter of research.

Unlabelled training set: In this proposal, the classifier is trained using a set of links collected automatically. The system analyses them and gives the user a list of prototypes representing concepts, while the user is only responsible for defining his or her interest, by labeling and picking one or more prototypes. Users intervention is unavoidable, given that the relevancy criteria depends solely on them. Recorders provide efficient crawlers that retrieve only relevant pages, but they depend entirely on the user.

As a result, we designed an efficient crawler, able to access web pages automatically, while requiring as little intervention as possible from the users. Some preliminary results of our implementation of the link classifier are shown in [11].

References

1. Aggarwal, C.C., Al-Garawi, F., Yu, P.S.: On the design of a learning crawler for topical resource discovery. ACM Trans. Inf. Syst. 19(3), 286–309 (2001)
2. Álvarez, M., Raposo, J., Pan, A., Cacheda, F., Bellas, F., Carneiro, V.: Crawling the content hidden behind web forms. In: ICCSA (2), pp. 322–333 (2007)
3. Anupam, V., Freire, J., Kumar, B., Lieuwen, D.F.: Automating web navigation with the webvcr. Computer Networks 33(1-6), 503–517 (2000)

4. Blythe, J., Kapoor, D., Knoblock, C.A., Lerman, K., Minton, S.: Information integration for the masses. J. UCS 14(11), 1811–1837 (2008)
5. Chakrabarti, S.: Focused web crawling. In: Encyclopedia of Database Systems, pp. 1147–1155 (2009)
6. Davulcu, H., Freire, J., Kifer, M., Ramakrishnan, I.V.: A layered architecture for querying dynamic web content. In: SIGMOD Conference, pp. 491–502 (1999)
7. de Assis, G.T., Laender, A.H.F., Gonçalves, M.A., da Silva, A.S.: Exploiting genre in focused crawling. In: String Processing and Information Retrieval, pp. 62–73 (2007)
8. Edwards, J., McCurley, K.S., Tomlin, J.A.: An adaptive model for optimizing performance of an incremental web crawler. In: WWW, pp. 106–113 (2001)
9. Fowler, H.W., Fowler, F.G.: Concise Oxford English Dictionary, 11th edn. revised. Oxford University Press, Oxford (2008)
10. He, H., Meng, W., Lu, Y., Yu, C.T., Wu, Z.: Towards deeper understanding of the search interfaces of the deep web. In: WWW, pp. 133–155 (2007)
11. Hernández, I.: Relc demo (2011), http://www.tdg-seville.info/inmahernandez/Thesis+Demo,
12. Lage, J.P., da Silva, A.S., Golgher, P.B., Laender, A.H.F.: Automatic generation of agents for collecting hidden web pages for data extraction. Data Knowl. Eng. 49(2), 177–196 (2004)
13. Liddle, S.W., Embley, D.W., Scott, D.T., Yau, S.H.: Extracting data behind web forms. In: ER (Workshops), pp. 402–413 (2002)
14. Mukherjea, S.: Discovering and analyzing world wide web collections. Knowl. Inf. Syst. 6(2), 230–241 (2004)
15. Pan, A., Raposo, J., Álvarez, M., Hidalgo, J., Viña, Á.: Semi-automatic wrapper generation for commercial web sources. In: Engineering Information Systems in the Internet Context, pp. 265–283 (2002)
16. Pant, G., Srinivasan, P.: Link contexts in classifier-guided topical crawlers. IEEE Trans. Knowl. Data Eng. 18(1), 107–122 (2006)
17. Partalas, I., Paliouras, G., Vlahavas, I.P.: Reinforcement learning with classifier selection for focused crawling. In: European Conference on Artificial Intelligence, pp. 759–760 (2008)
18. Raghavan, S., Garcia-Molina, H.: Crawling the hidden web. In: World Wide Web Conference Series (2001)
19. Shu, L., Meng, W., He, H., Yu, C.T.: Querying capability modeling and construction of deep web sources. In: Web Information Systems Engineering, pp. 13–25 (2007)
20. Vidal, M.L.A., da Silva, A.S., de Moura, E.S., Cavalcanti, J.M.B.: Structure-based crawling in the hidden web. J. UCS 14(11), 1857–1876 (2008)
21. Wang, Y., Hornung, T.: Deep web navigation by example. In: BIS (Workshops), pp. 131–140 (2008)

Web Trace Duplication Detection Based on Context

Chang Gao[1], Xiaoguang Hong[1,2], Zhaohui Peng[1], and Hongda Chen[1]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] Shandong Dareway Software Co., Ltd., Jinan, China
Gaochang26@163.com, {hxg,pzh}@sdu.edu.cn, sduchd@yahoo.com.cn

Abstract. Data Integration becomes more and more important with the rapidly spread of the internet and the study on entity trace becomes more and more important as a part of it. The entity trace is mainly extracted from the text fragments. There will be much duplication in the records because of the large scale, strong autonomy and the high redundancy features of the web sources. The processing of this problem often carries semantic features, which results in that the traditional integration method cannot be applied on it directly. In this paper, we propose a web trace duplication detection method based on unsupervised learning and context. We address the problem above by a new process on computing the comparison vector between two records based on the context, then acquiring the sample data automatically, training the classifiers with the sample data, and finally classifying the records.

Keywords: semantic features, web trace, duplication detection, unsupervised learning.

1 Introduction

As the rapidly spread of the internet, the web has drawn more and more attention for its vast amount of information. Sufficient analysis, mining and processing on the vast amount of the data can access to affluent information and knowledge, which can be widely applied to Public Opinion Analysis, E-commerce, Market Intelligence Analysis, and so on.

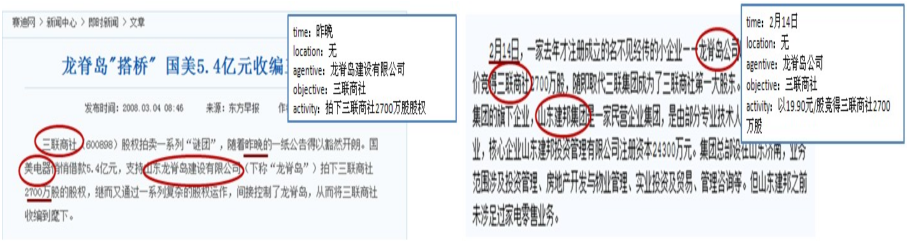
In the previous studies on web data integration facing analysis, they mainly deal with the more structured content in the web pages, but rarely pay attention on the unstructured text which contains deep information, such as the entity trace. This information plays a very important role in Market Intelligence Analysis System.

Entity trace is a certain activity of a certain entity happening at a certain time and a certain location [8]. Many entity traces can construct an enterprise's development trace. In Market Intelligence Analysis System, one enterprise's development trace is very important for applicants to refer to.

But web sources carry some special features, such as large scale, strong autonomy, and high redundancy; these features result in that (1) the same object has different expressions in different web sources, such as the named entity has full name, shortened form, alias, abbreviation; (2) the sentences with the same ideographic meaning will have different structures; (3) one entity trace may miss some attribute

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 292–301, 2011.

values or have some uncertain attribute values. For example, from Fig. 1(a), we may see that the time we extract from the text fragment is ambiguous and the location is missing. Comparing Fig. 1(a) and Fig. 1 (b), we may see that, the agentive has two different forms: Dragon's Back Island Construction Co., Ltd and Dragon's Back Island Company. And obviously, the entity traces extracted from the two text fragments are just representing one same trace.



赛迪网 > 新闻中心 > 即时新闻 > 文章

龙脊岛"搭桥" 国美5.4亿元收编

发布时间：2008.03.04 08:46 来源：东方早报 作

三联商社（600898）股权拍卖一系列"谜团"，随着昨晚的一纸公告得以豁然开朗。国美电器悄悄借款5.4亿元，支持山东龙脊岛建设有限公司（下称"龙脊岛"）拍下三联商社2700万股的股权，继而又通过一系列复杂的股权运作，间接控制了龙脊岛，从而将三联商社收编到麾下。

time：昨晚
location：无
agentive：龙脊岛建设有限公司
objective：三联商社
activity：拍下三联商社2700万股股权

(a)

2月14日，一家去年才注册成立的名不见经传的小企业——龙脊岛公司
价竞得三联商社2700万股，随即取代三联集团成为了三联商社第一大股东。
集团的旗下企业，山东建邦集团是一家民营企业集团，是由部分专业技术人
业，核心企业山东建邦投资管理有限公司注册资本24300万元。集团总部设在山东济南，业务范围涉及投资管理、房地产开发与物业管理、实业投资及贸易、管理咨询等。但山东建邦之前未涉足过家电零售业务。

time：2月14日
location：无
agentive：龙脊岛公司
objective：三联商社
activity：以19.90元/股竞得三联商社2700万股

(b)

Fig. 1. Two web traces can be extracted from the two text fragments separately. But some of their attributes may be missing or ambiguous. And we may see that the two records represent the same entity trace.

What we will do in this article is to put the entity traces who have the same ideographic meaning but different structures extracted from these unstructured text fragments together, a process that compares the records from different web sources and then determine which pairs of records represent the same entity trace, which is called web entity trace duplication detection.

In the whole data integration process, we extract many entity traces from many different web sources, these trace records are untreated. Some of them are imprecise, and some of them are incomplete. If we submit these records to the users to analyze directly, the result they reach will be imprecise, and even wrong because of the noises existing in the records, which is unacceptable for us. So, the data fusion before analysis becomes very significant. Data fusion aims at making up the missing attributes of the records, finding the true value and resolving the data conflicts, then merging the records representing the same entity trace to one record, finally providing complete, concise, and correct trace records for users. As the basis for data fusion, entity trace duplication detection can provide all the records that can be merged possibly, which can make the data fusion more simplified.

The traditional record duplication detection methods mainly deal with the records extracted from the structured part in the web pages. These records often appear in the form of label-attribute, and in the database, there is no semantic association between the different dimensions of a record. However, the entity traces are mainly extracted from the text segments, so there is a certain semantic relationship between a record's dimensions. Therefore, the traditional duplication detection methods cannot be applied to the entity traces duplication detection directly.

In this paper, we resolve the duplication detection problem in four steps: first, we divide the records in the database to several blocks; second, we will get the comparison vector of every two records in the same block, and this part is our emphasis that is different from others; the third step, we cluster the comparison vectors to three clusters – the matched cluster, the unmatched cluster and the possibly matched cluster; the last step, we utilize the matched cluster and the unmatched cluster as samples to train a SVM classifier, and then classify all the records.

The contribution of this paper is that we raise a process of calculating the comparison vector based on the context, because semantic relationship exists among the dimensions of a trace record, and they are affected by each other.

The rest of this paper is organized as follows. In Section 2, we will talk about the related work. In Section 3, we will give the ideas to compute the rough similarities and then make it precise by an iterative method based on the context. And we will show the other steps too. In section 4, the experiments are conducted, and section 5 will give our conclusion on the study.

2 Related Work

A lot of work has been proposed to solve this question. [1] uses the binary model to express each component x_i of the vector x. If it is matched on attribute i, then $x_i=1$, else, $x_i=0$. Expectation-maximization (EM) Algorithm is used to compute the probability p (x_i = 1|M). [2] uses the svmlight algorithm to learn how to combine the results of the comparison of more than one field in the records. [3] finds that the distance threshold of different duplicated records is different, and proposes an efficient algorithm to compute the different thresholds of the different entities. Ling Yan-yan [4] computes the duplication detection directly, she uses the VIPS algorithm to divide the page to several semantic blocks, and then computes the similarity of the record A and B by adding up the weights of all the similarities between each semantic block in A and B. The weight can be gained by means of iterative inequalities method. Kou Yue [5] proposes a deep web entity identification mechanism based on semantics and statistical analysis (SS-EIM), it is composed of text matching model, semantic analysis model, and group statistical model. By matching the text roughly, obtaining the representation relationship, and group statistical analysis, the author refines the recognition results using the text feature, semantic information, and constraint rules. Su Weifeng [6] proposes an Unsupervised Duplicate Detection (UDD) method, which is focusing on the duplication detection about instant enquiry. He treats all the two records in a single source as unmatched pairs. Then he trains a classifier with similarity weight on the basis of the unmatched pairs, which is then used to classify all the records to be matched between two query results, and concludes that all the records whose weight is greater than a threshold are matched. Next, he trains a SVM classifier using the matched and unmatched records. Finally, he re-trains the classifier with similarity weight by the SVM classifier until no new matched records are found. Huang Jian-bing [7] raises a conditional probability undirected graph model with a hidden variable to train the string edit distance among different fields. Based on this, he uses the SVM to combine the similarity weight of different fields, which can improve the accuracy of duplication detection.

At present, the study of duplication detection is mainly on the relational schema and the XML schema, which is in the initial stage on the deep web field.

3 Entity Trace Duplication Detection

3.1 Definition and Notation

Definition 1. Entity Trace: T= {time, location, agentive, objective, activity, context, URL, p} , where agentive is the initiator of the activity and objective is the entity affected by the activity, and both of them represent a real-world entity separately. Content is the text fragment where this trace is extracted from. URL is the page link where the content is from. P is the possibility of the reliability of this trace. [8]

Definition 2. Comparison Vector (CV) : CV (a, b) = (sim $(a.a_1, b.b_1)$, sim $(a, a_2, b.b_2)$…sim $(a.\ a_n, b.b_n)$), where a and b are records in the database, a_i and b_i are the dimensions of the record a and b separately. $0 \le \mathrm{sim}(a.a_i, b.b_i) \le 1$.

Definition 3. Entity Trace Duplication: we call two records A and B duplicated if and only if their comparison vector CV (A, B) ∈M, unduplicated if and only if CV (A, B) ∈U, where M represents the matched cluster and U represents the unmatched cluster. In addition, a third set P, which is called possible matched, is introduced here.

Through abundant observation, we may see that not all the attribute value is explicit and the missing of the attribute value is frequent. So, we use the rule bellow to calculate the similarity:

$$\mathrm{Sim}\ (a, b) = \begin{cases} 0 \text{ if a or b is missing} \\ x \text{ while both a and b exist} \\ 1 \text{ if both a and b are missing} \end{cases} \tag{1}$$

3.2 Block

Due to that there will be a lot of entity traces in the database, if we compute the comparison vector between every two records directly, the time complexity will be very large, which will affect the efficiency. Therefore, we divide the records to blocks before we compute the vector to reduce the time complexity. This part is not this paper's emphasis and existed methods like [9] are good enough to meet our needs.

3.3 Comparison Vector

In this paper, the vector can be composed of four components where we regard agentive and objective as the named entity. The four components are time, location, the named entity and activity. First, we need to calculate the similarities of the four components separately. Because the components' semantic role and their value's data type are different, we need to employ different methods separately. Second, we use them to compute the similarity of the text fragment by giving them different weights. Then, we compute the similarity of the four components using the similarity of the text fragment in return. This is an iterative process. We stop it until the range of the text fragment similarity is less than a threshold.

On the time dimension, we may get some uncertain value, such as yesterday night, several days ago. In this case, we desert this value and regard it as a missing case, then compute it using "(1)". On the other conditions, we firstly change the time format to a uniform format, and then compute the similarity according to the detail size of the expression of the time.

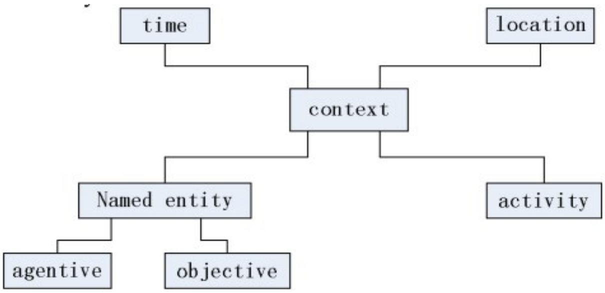


Fig. 2. The frame of the computing of the comparison vector

On the location dimension, we can create a location dictionary, using which we can calculate the similarity. For the cases of value missing, we can use “(1)”.

We can get a conclusion easily that the named entity (including the agentive and the objective attribute) and the activity are the two most important dimensions, and they can almost determine whether two records represent the same one trace.

Here, because that not every trace contains an objective, we consider to combine the agentive and objective together to the named entity. The similarity of the named entity is affected by the similarities of the agentive and the objective, but they have different weights (Fig. 2). We define it as follow:

$$\mathrm{sim}(\mathrm{entity}_1, \mathrm{entity}_2) = \alpha * \mathrm{sim}(\mathrm{agent}_1, \mathrm{agent}_2) + \beta * \mathrm{sim}(\mathrm{object}_1, \mathrm{object}_2) \quad (2)$$

Where $\alpha + \beta = 1$. [5] proposes a deep web entity identification mechanism called SS-EIM, which is an assistant we can use to compute the similarity.

Activity, which is the main component of a trace, is the most important dimension. The synonymous action verbs in the activity are the main reason in causing the duplication problem, such as “purchase” and “buy”. We can use word breaker to handle the activity sentence first, and then calculate a rough similarity of the activity based on the WordNet [10] and HowNet [11] methods.

After getting these similarities, we can begin to create the iteration. In the initial stage, we assume that the similarity of the context is 1. In the next step, we can compute the similarity of the context with the four similarities and their weights because the importance of them are different, the more important, the bigger the weight will be. We can compute it by the follow formula:

$$\mathrm{s}(\mathrm{c}_1, \mathrm{c}_2) = \mathrm{a} * \mathrm{s}_\mathrm{t} + \mathrm{b} * \mathrm{s}_\mathrm{l} + \mathrm{c} * \mathrm{s}_\mathrm{e} + \mathrm{d} * \mathrm{s}_\mathrm{A} \quad (3)$$

$s(c_1, c_2)$ is the similarity of the two contexts, s_t is the similarity of time, s_l is the similarity of location, s_e is the similarity of the named entity, s_A is the similarity of the activity and a, b, c, d is the weight of time, location, the named entity and the activity separately. a + b + c + d =1.

In the common case, if the similarity of the agentive and the objective is high, the activities of the two traces may represent the same one in all probability, in other words, the similarity of the activity can be affected by the agentive and the objective. So, we give the below formula:

$$\mathrm{s}_\mathrm{A} = \left(\mathrm{m} * \mathrm{s}_\mathrm{e} + \mathrm{n} * \mathrm{s}(\mathrm{c}_1, \mathrm{c}_2)\right) * \mathrm{s}_\mathrm{A} \quad (4)$$

Now, we give the whole comparison function below:

Algorithm 1. COMPARISON (A, B) return CV

```
Input: entity trace A and B
Output: the comparison vector of A and B
Procedure
[1] //initialize the similarity of the contexts with 1
  1:= the similarity of the contexts
[2] $S_t$ := the similarity of time
   $S_l$ := the similarity of location
       $S_a$ := the similarity of agentive
   $S_o$:= the similarity of objective
       $S_e$:= $\alpha*S_a+(1-\alpha)*S_o$ //the similarity of the named entity
[3] 1:= $\Delta S_c$ //initialize the range of the context similarity
[4] while ($\Delta S_c$>m )  //m is a threshold
[5]     $S_c$:=$a*S_t + b*S_l + c*S_e + d*S_A$
[6]     $S_A$:=$(m*S_e + n*S_c)*S_A$
[7]     compute $\Delta S_c$
[8] end while
[9] output the comparison vector CV
End procedure.
```

From the formula (3) and formula (4), we may see that in the iterative process, S_c is decreasing and S_A is decreasing affected by S_c. In return, S_A affects S_c to decrease. This process will continue until the range of the text fragment similarity is less than a threshold.

3.4 Cluster

In the paper [12], two regular patterns are raised: (1) if two records to be compared represent the same entity, the probability of high similarity on most of the dimensions will be high, and the similarities will not be high on every dimension if they represent two different entities. (2) If two records to be compared represent the different entities, the probability of low similarity on most of the dimensions will be high, and the similarities will not be low on every dimension if they represent the same entity. The two rules above are also applied to our problem, based on which we can choose the high probabilistic matched cluster and unmatched cluster as the samples of the next step automatically.

On this step, we use K-means algorithm to cluster the comparison vectors to three clusters: the matcher cluster, the unmatched cluster and the possibly matched cluster. K-means is a classic cluster algorithm. First, it chooses k centers, and then alters the k centers on the basis of the distance of the similarity iteratively. Finally, it stops until there is no alteration or some conditions are satisfied. In this paper, we can set k=3 and initialize the center with matched M, unmatched U and possibly matched P where P_i=0.5. The vectors closed to m belong to the matched cluster X_M and the vectors closed to U belong to the unmatched cluster X_U.

Here, because of the different weight of the dimensions, we modify the method of choosing the center:

$$\text{center}_k = \frac{1}{n_k}(a * \sum S_{ti}, b * \sum S_{li}, c * \sum S_{ei}, d * \sum S_{Ai}) \tag{5}$$

Where (S_{ti}, S_{li}, S_{ei}, S_{Ai}) ∈ $cluster_k$, and a, b, c, d are weights. a + b + c + d=1.

The cluster algorithm is below:

Algorithm 2. CLUSTER (the set of CV) return X_M , X_U and X_P

```
  Input: the set of comparison vectors
  Output: $X_M$ , $X_U$ , $X_P$
  Procedure
  [1] Choose M, U and P as the cluster centers
  [2] Classify the vectors to the cluster whose cluster center is the nearest
  [3] Use formula (5), recalculate the cluster centers
  [4] Repeat [2] [3] until there is no alteration or some conditions are satisfied
  [5] Output $X_M$, $X_U$, $X_P$
  End procedure.
```

3.5 Train the SVM

SVM is the support vector machine, which is a classification method based on supervise learning. It needs a trained sample to train the sorter, and then gain a classification model, finally classifies the test set with the model.

Before the last step, we have already had the matched cluster sample and the unmatched cluster sample. To make the result more precise, we use the SVM iteratively. First, we set the matched sample and the unmatched sample as input to train an initial SVM model, and then classify the rest comparison vectors, and put the most possibly correct ones in the X_M and X_U cluster separately. Then we use the new clusters to train a new SVM model. Repeat the process until there is no comparison vector to be classified. The comparison vectors belong to the last cluster X_M is the matched ones.

We describe the process as algorithm 3:

Algorithm 3. TRAIN_SVM (X_M, X_U, X)

```
Input: the set of comparison vectors X, the matched sample
       $X_M$, the unmatched sample $X_U$
Output: the final matched cluster $X_M$ and the final unmatched
       cluster $X_U$
  Procedure:
  [1] $T_M$:=$X_M$  $T_U$:=$X_U$
  [2] $X_N$ := X-($X_M$∪$X_U$)   //$X_N$ is the set to be classified
  [3] $svm_0$:=svm_train($T_M$, $T_U$)
  [4] k:=0
  [5] while $X_N$ is not empty
  [6]   ($W_M$, $W_U$):= svm_classify ($svm_k$, $X_N$)
  [7]   $T_M$ := $T_M$∪$W_M$
  [8]   $T_U$ := $T_U$∪$W_U$
  [9]   k++
  [10]  $svm_k$ := svm_train($T_M$, $T_U$)
  [11]  $X_N$ := $X_N$-($X_M$∪$X_U$)
```

```
end while
output $X_M$, $X_U$
End procedure
```

4 Experiment

This section is about our experiment study, and the main purpose is to find the best weights distribution when we compute the comparison vectors by experiments. Then we will compare our study with the traditional method.

In section A, we want to find a better allocation plan on how to distribute the weights by experiment. We will set different combinations and observe the result. In section B, we will use the weights we get in section A to compare our method with the traditional methods.

Here, we use A to represent the numbers of the duplicated records we recognize, use B to represent the number of the correct ones in the duplicated records we recognize, and use C to represent the number of the duplicated records that we do not recognize.

The definition of Recall and Precision is below:

$$Precision = \frac{\mathrm{B}}{\mathrm{A}} \qquad Recall = \frac{\mathrm{B}}{\mathrm{B+C}} \tag{5}$$

Where Precision represents the reliability of the result and Recall represents the cover of the correct result.

4.1 Section A

During the computing of the comparison vectors, we can analyze that the importance of the different dimensions is different when we decide whether two traces are the same one. So, we should give them different weights, and the more important the dimension is, the high weight it will have. We can get a better allocation plan by experiment.

We first fix the weights of time and location with different values, and we distribute the rest weights on the named entity and activity. (0.1,0.1,*,*) means that the weight of time and location are both be fixed to 0.1 and we will distribute the rest 0.8 weights on the named entity and activity.

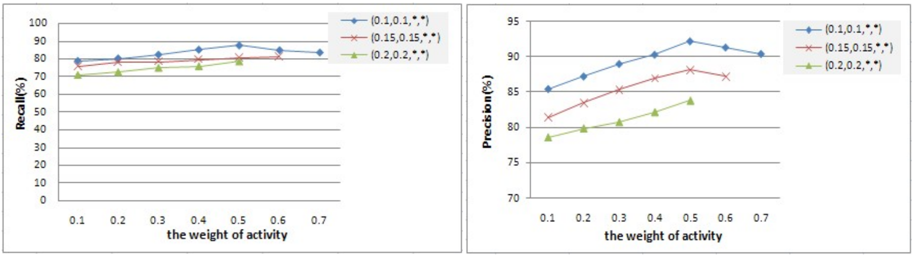


Fig. 3 and 4. They show the recall and the precision when we set different weights on the four dimensions

From Fig. 3 and 4, we may see that when we give high weight on the named entity and activity, the results of precision and recall are better. But too high weight will reduce the precision.

4.2 Section B

In this section, we will compare our method with the traditional methods. In traditional methods, the comparison vector is computed on the dimensions separately. There is no relationship between each other. In our approach, we should consider the semantic relationship between the dimensions besides the similarities of the dimensions themselves.

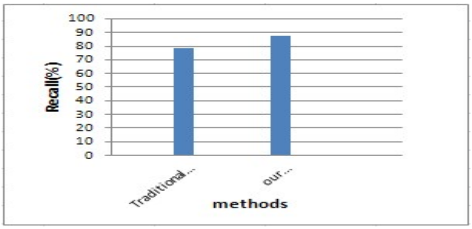


Fig. 5. The comparison on recall of the traditional method and our method

From Fig. 5, we may see obviously that the traditional methods cannot be applied to the duplication detection of the entity trace. The consideration of the semantic relationship among the dimensions is very important.

5 Conclusion

Along with the improvement of the importance of the internet, data integration is increasingly easier to draw our attention. The entity trace is also becoming more attractive as a component of the data integration. These traces are always hidden in the unstructured text fragments in the web pages, which results in that the relation of a trace's dimensions is not independent. When we compute the comparison vectors, we should also consider the semantic relation. Therefore, we should make some modifications before we use the traditional methods directly. In this paper, we raise a method to compute the comparison vector based on the context and the experiments show that it is more effective.

Acknowledgments. This work was supported in part by the National Key Technologies R&D Program No.2009BAH44B05, the National Natural Science Foundation of China under Grant No.61003051, the Natural Science Foundation of Shandong Province of China under Grant No.2009ZRB019RW, and the Key Technology R&D Program of Shandong Province under Grant No.2010GGX10108.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B(39), 1–38 (1977)
2. Bilenko, M., Mooney, R., Cohen, W., Ravikumar, P., Fienberg, S.: Adaptive Name Matching in Information Integration. IEEE Intelligent Systems 18(5), 16–23 (2003)
3. Chaudhuri, S., Ganti, V., Motwani, R.: Robust Identification of Fuzzy Duplicates. In: Proceedings of the 21st International Conference on Data Engineering, Washington, DC, USA, pp. 865–876 (2005)
4. Ling, Y.-y., Liu, W., Wang, Z.-y.: Entity Identification for Deep Web Data Integration. Journal of Computer Research and Development 43, 46–53 (2006)
5. Kou, Y., Shen, D.-R., Li, D., Nie, T.-Z.: A Deep Web Entity Identification Mechanism Based on Semantics and Statistical Analysis. 19(2), 194–208 (2008)
6. Su, W., Wang, J., Lochovsky, F.: Record Matching over Query Results from Multiple Web Databases. IEEE Transtraction on Knowledge and Data Engineering 22(4), 578–589 (2010)
7. Huang, J.-b., Ji, H.-b.: An adaptive similarity learning approach to record linkage. Journal of XIDIAN University 34(2), 126–130 (2007)
8. Yao, C.: Research on the extraction of Web entities and discovery of entity activities. In: Partial Fulfillment of the Requirement for the Degree of Doctoral of Philosophy, Peking University (2008)
9. de, V.T., Ke, H., Chawla, S., Christen, P.: Robust record linkage blocking using suffix arrays. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, New York, NY, USA, pp. 305–314 (2009)
10. Agirre, E., de Lacalle, O.L., Fellbaum, C.: SemEval-2010 task 17: All-words word sense disambiguation on a specific domain. In: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval 2010, pp. 75–80 (2010)
11. Wu, J., Wu, Z.-h., Li, Y.: Web Service Discovery Based on Ontology and Similarity of Words. Chinese Journal of Computers 28(4), 595–602 (2005)
12. Dong, Y.: Research on Key Issues in Deep Web Data Integration. In: Partial Fulfillment of the Requirement for the Degree of Doctoral Of Philosophy. Shandong University (2010)

A Framework for Incremental Deep Web Crawler Based on URL Classification

Zhixiao Zhang[1], Guoqing Dong[1,2], Zhaohui Peng[1], and Zhongmin Yan[1]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] Shandong Dareway Software Co., Ltd., Jinan, China
tyzcyx@yeah.net, {dgq,pzh,yzm}@sdu.edu.cn

Abstract. With the Web grows rapidly, more and more data become available in the Deep Web · But users have to key in a set of keywords in order to access the pages from some web sites. Traditional search engines only index and retrieve Surface Web pages through static URL links, because Deep Web pages are hidden behind the forms. However, the amount of information contained in the Deep web is not only far more than the Surface Web, the information of Deep Web is more valuable than the Surface Web. As Deep Web Pages change rapidly, how to maintain the Deep Web pages which were crawled fresh and to crawl the new Deep Web pages is a challenge. A framework for incremental Deep Web crawler based on URL classification is proposed. According to the list page and leaf page, the URL that is related with the page can be divided into two parts: list URL and leaf URL. The framework not only crawls the latest Deep Web pages according to the change frequency of list page, but also crawl the leaf pages which often change.

Keywords: Deep Web, Incremental Crawl, URL Classification.

1 Introduction

The World Wide Web include Deep Web and Surface Web, Surface Web is defined as follows: traditional search engines can index the pages in the Surface Web through hyperlink, the pages mainly composed of static pages. However, the Deep Web is qualitatively different from the surface Web. Deep Web sources store their content in searchable databases which only create dynamic contents in response to a user request. According to the recent study, 50% of Deep Web is domain-oriented.

The Deep Web cover almost all areas which includes Education, economy, sports etc, have a large amount of information and contain high valuable information. But these information resources are hidden behind query form, so there should be a special crawler to crawl the Deep Web pages. It should be able to download query forms, fill them, submit them by users and crawl the related page according to the URL [1] that is created by the form submitted. The general architecture of Deep Web crawler is shown in fig1.In order to save time and resources, the Deep Web crawler should crawl new pages and those pages which often change in the content [2]. This paper proposes a novel framework to incremental crawl Deep Web pages using URL classification. It crawl those pages whose contents are latest in the Web and try it best to update those

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 302–310, 2011.

pages which have been crawled and stored in the local repository by re-crawling the pages[3],[4],[5]. The above method can reduce the times of crawling the unchanged pages. In this paper, the contents are organized as follows: the second section describes related work which has been carried out in this area; the third section describes how the framework for incremental Deep Web crawler based on URL classification to work; the fourth section describes the performance of the framework; the last section concludes proposed work.

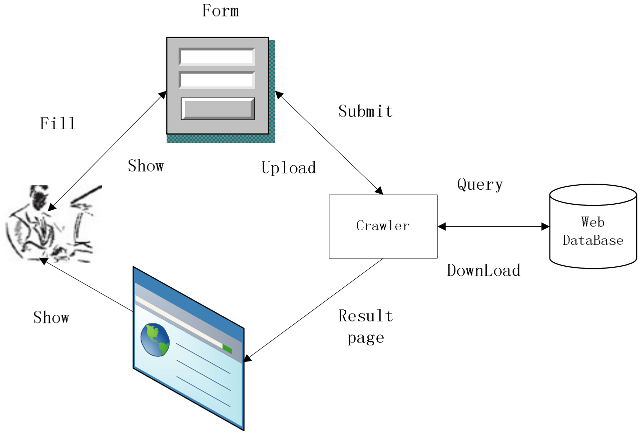


Fig. 1. Deep Web crawler crawl the Deep Web pages

2 Related Work

Nowadays, more and more researchers work at the field of crawling the Deep Web, and there are lots of related works associated with crawling technology.

At present, the Hidden Web Exposer (HIWE) [6] leading by Stanford University is the most classic prototype system. This project designs and develops a crawler which can extract the contents of Deep Web. In this system, crawler's manager takes charge of the searching process. It analyzes the Web page then delivers the page which containing the forms to the form processor. The form processor extracts forms from the pages first, then fills the forms automatically according to information from prepared data set, and then submits the synthetic URLs to crawler's manager for downloading the related pages. In view of the need for filling the forms automatically, this system need the users prepare the corresponding form data sets first. HIWE can only be used for a particular area, and must be completed in artificial help.

ShopBot which developed by Washington university uses heuristic methods in specific field to fill out a form in order to contrast the goods in this field. The operation can be divided into two stages: learning-offline stage and online product comparison stage. In learning stage, ShopBot determines how to fill forms and how to analyze the result page, so that, it gets the model information for this site. In the stage of comparison, ShopBot extracts result information using the model information created by the first stage, then looks for the price optimal product which must meet user requirements .We can see that its narrow research fields does not apply to large-scale integrated information.

Panagiotis G Ipeirotis in the research team of Columbia University [7] work at finding a method to classify automatically by backend database connected. first it create a set of rule-based Classifier (Classifier) using the technology of machine learning ,then it transformed classifier into the URL and query the backend database by these URLs, Finally ,classify the data according to the query result, but only for text database classification.

However, despite the crawler above can crawl the Deep Web pages, but it do not have an explicit way for crawling the page incrementally. Obviously unchanged web pages not need to crawl again and it's important to find intervals between the page changes.

3 The Architecture of an Incremental Deep Web Crawler

This section proposes the architecture of an Incremental Deep Web Crawler and describes how this crawler to efficiently incremental crawl Deep Web Pages in detail. Deep Web pages can be divided into the list page and the leaf page according to the content of the pages, so the URL that corresponds with the page also can be divided into list URL and leaf URL. The list page refer to the page which composes of a number of different data records from Web database, but the leaf page refer to the page which composes of the details of a data record in the list page. The list page can connect to a certain leaf page through the URL-link that corresponds with one data record. Most of information which needs to be extracted and analyzed is mainly from the leaf page, it requires separately dealing with the leaf page. This crawler makes a prediction about

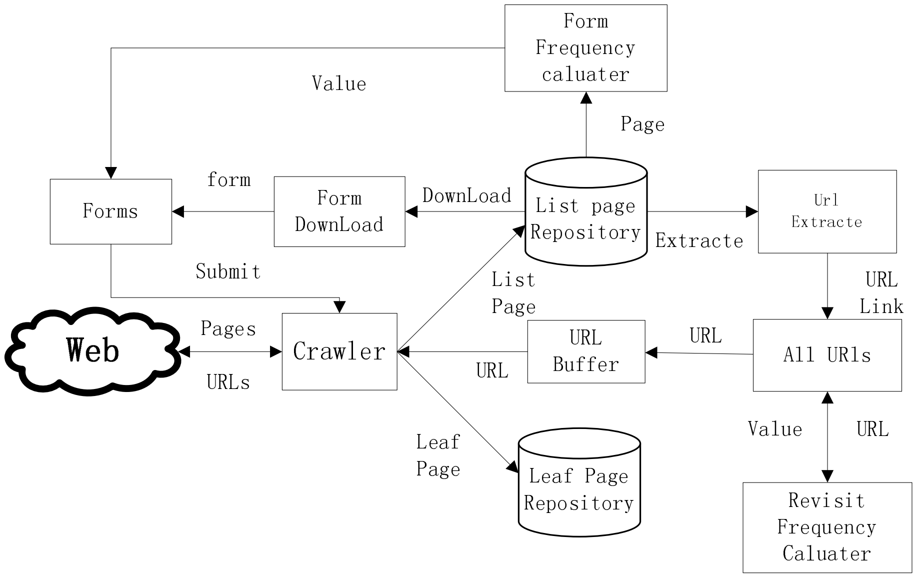


Fig. 2. The architecture of an Incremental Deep Web Crawler

the frequency of submitting query form according to the change frequency of the list page which is created by the same type of query form and resubmits the same type of query form to crawl the list page at a certain time interval. The crawler makes a prediction about the change frequency of leaf page and re-crawl the leaf page. The crawler starts work with the query form initialization; submits the query form to the crawler. The crawler will do different things according to the URL that is created. If the URL belongs to the list URL, when the list page is crawled, there will be three steps, first, the calculator of change frequency of list page calculates the change frequency of list page which belongs to the same type of query form, so the frequency of submitting the kind of query form will be obtained. Second, if the list page contains the form that is different from the query form of form repository and the above form belongs to query form, the form will be download and stored in the form repository. Third, The URL extractor will only extract the leaf URL from the list page according to certain rules, then the leaf URL will compares with the URL of AllURLs, if the same URL exists, the corresponding leaf page will be checked to see whether there have been changes, if the leaf page changes, the page will be re-crawled, Otherwise, the leaf page will be abandoned. Back to top, if the URL belongs to the leaf URL, then the crawler will crawl the leaf page, store it in the leaf repository and record the crawling time. The crawler will adjust strategy itself. It deals with the URL that is created by the query form with the ability of the N%, and the remaining deals with the leaf URL. These two abilities will change to make the crawler continue working.

The architecture of the Deep Web crawler includes the following modules:

Deep Web crawler module

Query form extractors
The calculator of change frequency of list page
URL extractor
The calculator of change frequency of leaf page
Leaf URLs Buffer

3.1 Deep Web Crawler Module

Deep Web crawler is a kind of search engine crawling the Deep Web pages. What it deals with is a six-tuple. The six-tuple is defined as follows:

{type, URL, urlfresh, crawltime, formtype}.
type: the type is divided list URL and leaf URL in all.
URL: the URL that is created by the query form or the leaf URL that correspond with the leaf page.
urlfresh: it is used to logo page. ”new” Page includes the pages that never appear until now and the change pages. ”old” page is the unchanged pages.
crawltime: when page is crawled is considered the crawltime.
formtype: formtype is a static n-tuple, it is compose of n different form items. Each form item includes one label name, label type, value. It is defined as follows:{label name, label type, value}.Different formtype includes different form item, the same formtype not only includes the same numbers of form item, but also the corresponding form item is the same, of course the order of the form item doesn’t matter. Take looking for job for example,

Table 1. formtype1

label name	label type	value
education	select	undergraduate
workplace	select	Beijing
keywords	input	java

Table 2. formtype2

label name	label type	value
workplace	select	Beijing
education	select	undergraduate
keywords	input	java

Table 3. formtype3

label name	label type	value
education	select	undergraduate
salary	select	5000
keywords	input	java

Table 4. formtype4

label name	label type	value
education	select	undergraduate
workplace	select	Beijing
keywords	input	java
Work time	checkbox	8 hours

formtype1=formtype2, the remaining are different.

The crawler crawl the Deep Web pages according to URL. If the type of URL is list URL, when the list page is crawled, the change frequency of the list page which belongs to the same type of form will be calculated. If the type of URL is leaf URL and urlfresh is new, then crawl the leaf page and record the crawltime. At the same time, calculate the time interval between the leaf page that is just crawled and the leaf page that has been crawled.

3.2 Query Form Extractors

Form Extractor is a rule-base extractor. It mainly extracts query form. Because there are some login forms, Mail forms in the Deep Web pages, these forms' existence bring difficulty in extracting query form. In order to solve this problem, three rules are proposed:

There is a form s, if the s contains the element tag name which is related with the user name, password etc, s will be ignored.

There is a form s, the element attribute of s is “input”, but the number of non-button field is less than 2,s will be ignored.

There is a form s, if s only contains a checkbox or a select list, and s is a limited easy form, s will be ignored.

Rule-a mainly removes the login form, registration form, purchase form; Rule-b mainly removes the forms of common search engines; Rule-c mainly removes easy and limited form. As formtype is composed of different form items, when a form is extracted, the form will be divided to get form items, then classify according to label name and label type, at last integer a wide coverage search form. Because Deep Web is field-oriented, the same topic of search form commonly shares some same or similar pattern structure. For example, a job search form commonly is designed as follows: first is search classification, like keyword search. Then advanced search, it includes some form items like functional category, release date▯ work place. In the integrated query form, filling a form more accurate will reduce the times of crawl the unchanged pages.

3.3 The Calculator of Change Frequency of List Page

The frequency of submitting search frequency is obtained by calculating the change frequency of list page which belongs to the same type of search form. The calculating method will be described in the follows. The URL is a six-tuple, after filling the form and submitting form, the crawl can recognize the formtype and the type of the URL. After crawling the corresponding list page and store in the list page repository, this page will be compared with the list page which belongs to the same type of query form. Because the list page is shown as list data record, therefore, changes range of list page is based on the number of changed data records. Here set the threshold in the list page change range. When the list page just being crawled compare the list page of list page repository which belongs to the same type of query form, if the change range is greater than the threshold, then the list page is considered changed and record the crawl time in order to calculate the crawl time interval. If the change range is less than the threshold, the page is considered unchanged. For the changed list page , when the crawl time interval is calculated, the frequency of submitting the same type of query form will be achieve. For example, there is a form whose formtype is f, the list page just being crawled is p1, and the latest list page belonging to f in list page repository is p2. Compare the p1 with p2, if p1 changes, then the time interval of submitting f are the crawltime difference of p1 and p2.of course, the threshold is best set close to the number of list data record in the list page. Then the crawler will crawl the list page which includes all new leaf URL. The algorithm that associates with the frequency of submitting the same type of form is as follows:

Algorithm1: Step of an incremental crawler about the frequency of submitting the same type of form. We assume list page repository is not empty from the beginning. It concludes new URLs and old URLs

```
while (submitting time arrives)
Submitting form X
Page <— Crawl (URL)
if (Page∈ X and Page changed)then
        calculate (change frequency)
```

```
    achieve the re-submitting time
    submitting time <— re-submitting time
else
;
```

3.4 URL Extractor

URL Extractor extracts URLs which index leaf pages from certain list page, then compare with the URLs of All URLs, if the same URL exists in the All URLs, the urlfresh of the URL which is extracted is assigned old. If not, the urlfresh is assigned new, at the same time submit the URL to the URL Buffer.

3.5 The Calculator of Change Frequency of Leaf Page

This calculator mainly deals with the leaf pages whose urlfresh is old. When the URL is extracted, the URL is submitted to the URL Buffer until the related leaf page is crawled. After that, record the crawl time and compare the leaf page just being crawled with the leaf page of leaf page repository that has the same URL. Here set the threshold in the leaf page change range. Assume that the change range is greater than the threshold, the page is considered to have been changed, if not, the page is considered not to be changed. Base on the above suppose, if these two pages are the same, and then discard the leaf page just being crawled. If not, then update the leaf page, in fact, to replace the leaf page of repository with the leaf page just being crawled. At this moment, achieve the change interval by calculating the crawltime difference. If the leaf page changes frequently after that, the Leaf URL Buffer will send message to the AllURLs to get the leaf URL in a certain time interval.

3.6 Leaf URL Buffer

The leaf URL Buffer is designed to increase the efficiency of the Deep Web crawler. In order to keep the URL Buffer not empty, When the number of the URL in the leaf URL Buffer reduced to close to one , the URL Buffer will send message to the AllURLs to get the new URL and the old URL satisfying crawling time interval. In above case, the crawler will never be at rest as a result of obtaining not URL from the URL Buffer. Of course, in the case of submitting form, the crawler will not be at rest. The algorithm that associates with leaf URL is as follows:

Algorithm 2: Step of an incremental crawler about leaf page .We assume AllURLs is not empty from the beginning. It concludes new urls and old urls.

```
while (true)
URL <— select(AllURLs)
if (urlfresh=new) then
    Page <— Crawl (URL)
else
    Page <— Crawl (URL)
    if(Page changed)then
        calculate (change frequency)
    else
        discard (Page)
```

4 Performance Evaluation

As the Deep Web pages dramatically increase, fresh or updated information retrieval has becomes a difficult problem for the traditionally search engines. With the size of Deep Web contents increasing, the above problem becomes more and more hard. The number of web pages which have gone under up-dation increases, with the increase in web size, the results decreased. The proposed incremental Deep Web crawler can crawl the new pages and update the Deep web pages that are already crawled and make the page repository fresh. To evaluate the performance of the proposed crawler, freshness of leaf page repository is considered an important measure. In order to evaluate the freshness, several experiments are done. First define the freshness [8] of leaf page repository at the time t:

$$F\ (D,\ t) = 1/N\Sigma F\ (e_i,\ t),$$

$$F(e_i, t) = \begin{cases} 1, \text{if page } e_i \text{ is up to date at the time t} \\ 0, \text{otherwise} \end{cases}$$

And N is the total number of leaf page belonging to the same type of query form of the leaf page repository.

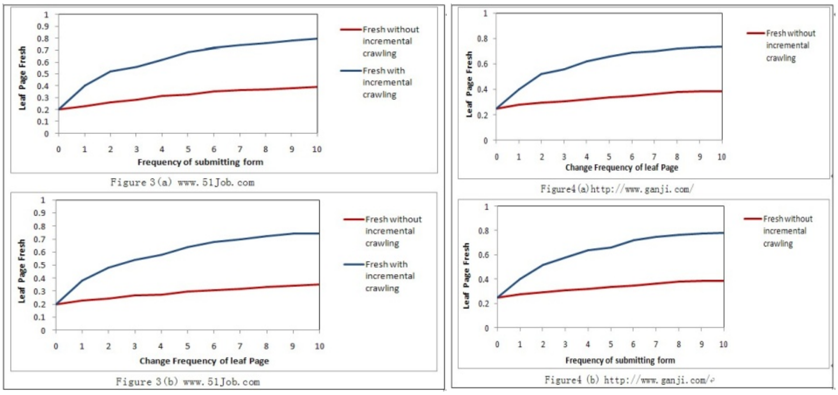


Figure 3(a) www.51Job.com

Figure4(a)http://www.ganji.com/

Figure 3(b) www.51Job.com

Figure4 (b) http://www.ganji.com/

Fig. 3. The difference between the fresh without incremental crawling and the fresh with incremental crawling

The frequency of submitting form and change frequency of leaf page are tested separately. Two Deep Web sites [9] www.51Job.com and http://www.ganji.com are selected as the crawling target website. The above figure shows the difference between the fresh without incremental crawling and the fresh with incremental crawling. The more fresh leaf pages are crawled by the proposed incremental Deep Web crawler.

5 Conclude

In this paper, an incremental Deep Web Crawler based on URL Classification has been designed. It not only crawls the new Deep Web pages but also keep the leaf page repository updated with the new and updated pages. It is based on calculating the time period between the two successive revisits for leaf pages and between the two successive resubmitting the same type of query form. As the future work, the architecture of Incremental Deep Web crawler which can crawl many different domains will be study. After that, the crawler can crawl efficiently Deep Web pages of different domains.

Acknowledgments. This work was supported by the National Key Technologies R&D Program No.2009BAH44B02, Key Technology R&D Program of Shandong Province under Grant No. 2010GGX10108.

References

[1] Cho, J., Garcia-Molina, H., Page, L.: Efficient crawling through URL ordering. In: Proceedings of the 7th World-Wide Web Conference (1998)

[2] Cho, J., Garcia-Molina, H.: Estimating frequency of change. Technical report, Stanford University (2000)

[3] Cho, J., Garcia-Molina, H.: The Evolution of the Web and Implications for an Incremental Crawler. In: Proceedings of the Twenty-Sixth VLDB Conference, Cairo, Egypt, pp. 200–209 (2000)

[4] Meng, T., Yan, H.F., Wang, J.: A model of efficient incremental spider for the Chinese Web and its implementation. Journal of Tsinghua University (Science and Technology) 45(S1), 1882–1886 (2005) (in Chinese with English abstract)

[5] Meng, T., Yan, H.F., Wang, J.M.: Web Evolution and Incremental Crawling. Journal of Software 17(5) (May 2006)

[6] Sharma, A.K., Gupta, J.P., Agarwal, D.P.: A novel approach towards management of Volatile Information. Journal of CSI 33(1), 18–27 (2003)

[7] Qprober Research Group (October 2005), acessible at http://qprober.CS.columbia.ed

[8] Cho, J., Garcia-Molina, H.: Synchronizing a database to improve freshness. In: Proceedings of the 2000 ACM SIGMOD (2000)

[9] Key Technology R&D Program of Shandong Province under Grant No. 2010GGX10108

[10] Chakrabarti, S., van den Berg, M., Dom, B.: Focused crawling new approach to topic-specific web resource discovery. In: Proceedings of the 8th World-Wide Web Conference (1999)

[11] Bhatia, K.K., Sharma, A.K.: A Framework for an Extensible Domain-specific Hidden Web Crawler (DSHWC). Communicated to IEEETKDE Journal (December 2008)

[12] Bhatia, K.K., Sharma, A.K.: A Framework for Domain-Specific Interface Mapper (DSIM). International Journal of Computer Science and Network Security, IJCSNS 2008 (2008)

[13] Dixit, A., Sharma, A.K.: Self Adjusting Refresh Time Based Architecture for Incremental Web Crawler. International Journal of Computer Science and Network Security (IJCSNS) 8(12) (December 2008)

[14] Cho, J., Roy, S.: Impact of Web search engines on page popularity. In: Proc. of the 13th World-Wide Web Conf., pp. 20–29. ACM Press, New York (2004)

Query Classification Based on Index Association Rule Expansion[*]

Xianghua Fu, Dongjian Chen, Xueping Guo, and Chao Wang

College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen Guangdong, 518060, China
fuxh@szu.edu.cn, winter_chen@msn.cn,
{GXPing1202,shendamrwang}@163.com

Abstract. Query classification can improve the query results of search engine, but the existing query classification methods which use extra web resources to enrich query features easily result in high delay. In this paper, a query classification based on index association rule expansion (IARE-QC) is proposed. IARE-QC uses an index based query classification framework to reduce the response time through transforming the query classification problem in online phase to the equivalent index term classification in offline phase. Moreover, in order to get more accurate feature enrichment of index term, we propose a novel algorithm which called index association expansion based on similarity voting (IARE-SV) to determine the category labels of index term. The experiment results on the search engine simulation environment show that IARE-SV can get much better query classification performance than the common simple voting (SV) method.

Keywords: query classification, query expansion, association rule analysis.

1 Introduction

Because we usually can not get satisfactory query results from search engines with the simple query words matching, researchers propose the idea of query classification which map the queries submitted by users to a pre-specified target category through a certain method [1-3]. The existing query classification methods usually use extra web resources, such as the search results of the query or other open taxonomies, to enrich the query and create the target taxonomy [4]. For example, Broder[5] proposes a method to classify the query terms into the navigation, information and transaction category. Dou Shen[6, 7] enrich query by the search results, and classify the query terms into the intermediate ODP Directory by SVM, then map the intermediate taxonomy into KDDCUP's pre-defined 67 categories. These approaches have obtained good classification effects, however the shortcoming is that they will cost much time. So they are intolerable to the high speed response of search engine's request.

To solve the high delay problem, some improved methods are proposed. For example, Gabrilovich[8] analyzes the architecture of search engine and proposes an offline classification model. Grati[9] presents a classification model that can pre-calculate the

[*] This work is supported in part by National Natural Science Foundation of China (60903114, 61003271, 61001185), Natural Science Foundation of Guangdong Province (7301329), Scienceand Technology Foundation of Shenzhen City (JC201005280463A).

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 311–318, 2011.

features of query. Other studies such as Beitzel[10, 11] use the query logs as the training dataset to reduce the processing time and computing resource consumption.

In this paper, we propose a novel query classification method which is based on index association rules expansion (IARE-QC). Moreover, in order to get more accurate feature enrichment of index term, we propose a novel algorithm which called index association expansion based on similarity voting (IARE-SV) to determine the category labels of index term. Experimental results show that our method can get better query classification result.

2 Query Classification Model Based on Indexes

Search engines return search results with the keyword matching technology. When the search engine receives a query $Q=\{q_1,q_2,...,q_{|Q|}\}$, it finds the web pages $\mathcal{P}_i=\{p_1,p_2,...,p_{|\mathcal{P}_i|}\}$, which indexed by one query word q_i in the search engine at first, and then returns the intersection $\mathcal{P}_1 \cap \mathcal{P}_2 \cap ... \cap \mathcal{P}_{|Q|}$ as the query results. The search procedure essentially is to find the same index term *k* (*k*=*q*) in search engine's indexes with some query word *q*. So if we can classify the index term *k* in advance and tag the category labels to the index term *k*, then the classification result of the query word *q* can be obtained directly from *k*'s category labels when users input the query. According the above analysis, we propose a novel query classification model based on indexes, where the query classification is transformed to index term classification. The model has three distinct components, which are described as follows:

(1) To construct the text classifier for a giving target taxonomy.
(2) To obtain the text of the top *n* pages of every index term of the search engine index, and input them into the text classifier and get the classification result, then compute the category of the index term a voting strategy, and at last add the category labels to the index term.
(3) To search the category labels of corresponding index keyword in the search engine, and determine the category labels as the classification result of the query when the user inputs a query.

The component (1) and (2) are accomplished offline, and the online component (3) will be executed very quickly.

3 Index Terms Classification Based on Similarity Voting

Our main objective is to classify each index term and obtain their category labels. Because all the index terms are single word, we need to enrich their features expression before classifying them. The method to enrich features of an index term *k* is to choose full text information of the web pages which indexed by *k*. We only select the top *n* web pages which return from search engines with high page rank.

3.1 Index Term Classification

Given an index term k and a category label set $\mathcal{C}$, selecting the top *n* web pages indexed by *k* as a document set $\mathcal{D}_k=\{d_1,d_2,......,d_n\}$, the index term classification is a function $f_{\mathcal{D}_k 2\mathcal{C}}:\mathcal{D}_k \to \mathcal{C}$, which determine the category label c_k of k.

The function $f_{\mathcal{D}_k 2\mathcal{C}} : \mathcal{D}_k \to \mathcal{C}$ is usually implemented by calculating the conditional probability $P(c_i \,|k)$ at first and then selecting the category c_k with the largest conditional probability $c_{\max} = \arg\max_{c_j \in \mathcal{C}} P(c_k \,|k)$ [8, 12, 13]. Generally $P(c_j \,|k)$ can be calculated by Equation (1):

$$P(c_j \,|k) = \sum_{d_i \in \mathcal{D}_k} P(c_j \,|k, d_i) \cdot P(d_i \mid k) \tag{1}$$

Supposing $P(k \,|c_j, d_i) \approx P(k \mid d_i)$, Equation (1) can be transformed to Equation (2):

$$P(c_j \,|k) = \sum_{d_i \in \mathcal{D}_k} P(c_j \mid d_i) \cdot P(d_i \mid k) \tag{2}$$

Given a document d_i, the probability $P(c_j \mid d_i)$ can be calculated by the text classifier. Moreover, because d_i belongs $\mathcal{D}_k$ which indexed by *k*, we can simple set $P(d_i \mid k) = 1$. Then Equation (2) can be transformed to Equation (3):

$$P(c_j \,|k) = \sum_{d_i \in D_k} P(c_j \mid d_i) \tag{3}$$

Equation (3) is a relatively simple, so we call it the simple voting method (SV). In the simple voting method, each web page used in query enrichment is supposed to have the same weight that viz. $P(d_i \mid k) = 1$. Actually, the correlation between the index term and the enrichment web pages is different. So we propose a new voting method: voting based on similarity, this method calculates equation (2) with the similarity of the index term and the web pages.

3.2 Voting Based on Similarity

In order to calculate the similarity of the index term *k* and a web page d_i which is used to enrich the feature, we first try to expand the index keyword *k* to obtain an expanding feature vocabulary $\mathcal{E}_k = \{e_1, e_2, ..., e_{|\mathcal{E}_k|}\}$, and express *k* with $\mathcal{E}_k$, and then calculate the cosine similarity between $\mathcal{E}_k$ and each page in the set $\mathcal{D}_k$. The similarity value will act as voting weights to calculate $P(d_i \mid k)$ in equation (2).

The similarity of *k* and d_i is equal to the similarity of $\mathcal{E}_k$ and d_i, which can be calculated as Equation (4).

$$Sim(k, d_i) = Sim(\mathcal{E}_k, d_i) = \frac{\sum_{j=1}^{|\mathcal{E}_k|} w_j \times w_{ij}}{\sqrt{\sum_{j=1}^{|\mathcal{E}_k|} w_j \times \sum_{j=1}^{|\mathcal{E}_k|} w_{ij}}} \tag{4}$$

Where $\left\{w_1, w_2, ..., w_{|\mathcal{E}_k|}\right\}$ is the weight vector of $\mathcal{E}_k$, w_{ij} is the weight of each word in d_i. With $Sim(k, d_i)$ as conditional probability, Equation (5) can be derived the from equation (2) as follows.

$$P(c_j|k) = \sum_{d_i \in \mathcal{D}_k} P(c_j \mid d_i) \cdot Sim(k, d_i) \tag{5}$$

Now, the key problem is how to expand k with $\mathcal{D}_k$ in Equation (5).

4 Index Term Expansion Based on Association Support

The expansion of the index term $k \to \mathcal{E}_k = \{e_1, e_2, ..., e_{|\mathcal{E}_k|}\}$ is similar to the query expansion. We consider implementing it with the association rule mining of words [14]. In order to getting the association support of the index term and the expanding feature word on the search engine's index, here we propose an index term expansion based on the association support.

Supposing each document in $\mathcal{D}_k$ has a specific ID, and $\mathcal{V}$ is the feature word set of $\mathcal{D}_k$, then the inverted index of $\mathcal{D}_k$ can be denoted as $(\mathcal{V}, \mathcal{I})$. $\mathcal{I} = \left\{ind(t_1), ind(t_2), ..., ind(t_{|\mathcal{V}|})\right\}$ is the index table, where $ind(t)$ contains a document ID list of documents that having word t ($t \in \mathcal{V}$), and other related information. $ind(t) = \left(\langle ID_{l,1}, w_{t_{l,1}}\rangle, \langle ID_{l,2}, w_{t_{l,2}}\rangle, \ldots\right)$, where $ID_{l,1}$ is the ID of the first document which have feature word t , and $w_{t_{l,1}}$ is the weight of t in the document which ID is $ID_{l,1}$.

The association support of index term k and the feature word t can be calculated according the invert index. Firstly, we take ID as a collection element to calculate $ind(t) \cap ind(k)$, and get the index items $\left(\langle ID_{l,1}, w_{t_{l,1}}\rangle, \langle ID_{l,2}, w_{k_{l,2}}\rangle, ...\right)$ in which $\left(ID_{l,1}, ID_{l,2}, ...\right)$ are both appeared in $ind(t)$ and $ind(k)$. Then we can calculate the association support between the index term k and the feature word t by Equation (6):

$$\text{support}(k,t) = \frac{1}{n} \sum_{ID_{l,j} \in ind(t) \cap ind(k)} w_{t_{l,j}} w_{k_{l,j}} \tag{6}$$

Equation (6) can be used to determine the expanding feature vocabulary $\mathcal{E}_k$ of the index term k . We also define an expansion parameter α, which is the percentage of the number of the expansion $\mathcal{E}_k$ with all the feature word of the index table. α is useful to control $|\mathcal{E}_k|$. In this paper we call the voting method based on Equation (5) and (6) as the Similarity Voting Based Index Association Rule Expansion (IARE-SV). Since the inverted index is a widely way of the index in information retrieval and search engines, the IARE-SV is hopeful to improve the query classification efficiency.

5 Experiment and Evaluation

We take some queries as the inputs of search engine and get their first 100 web pages. Then we establish the index of these queries to simulate the search engine's inverted index. The taxonomy of the portal site's news channels is chosen as the target taxonomy of query classification tasks. Moreover the news web pages of different channels from Tencent[1] are crawled to train the LIBSVM[2] classifier. The experiment dataset is shown in table 1, which includes the target categories, the number of web pages for training in each category, the number of queries before labeled for testing in each category, and the number of queries labeled for testing in each category. The queries for testing are selected from Baidu Search Billboard[3].

Table 1. Target category and the number of pages

Target category	The number of web pages for training	The number of queries before labeled for testing	The number of queries labeled for testing
car	2156	125	125
news	2890	40	87
sport	7453	30	41
finance	6132	55	56
game	2276	100	100
military	737	17	22
movie	1512	54	71
music	1962	36	44
IT	2984	218	218
TV	2017	62	68
variety show	1276	50	53
stars	2216	217	220

We designed two experiments with different voting method. The first is the simple voting method (SV) mentioned, the second is our method IARE-SV. The evaluation measures include the precision, recall and F1-measure. α is set as 0.3 in our experiments. And then we comprise experimental results of two methods of SV and IARE-SV. We chose the web pages of the 12 categories in Table 1 to train the LIBSVM classifier. All the query words having been labeled in Table 1 act as test data. To analyze the influence of the parameter n, we set n with 10, 20, 40, 60, 80, and 100.

Figure 1 is the average value of precision of SV and IARE-SV in all data. Figure 1 shows that with the increase of n, the precision of classification becomes higher. When n is 40, IARE-SV get the max value of 90%. SV get the max value of 77% when n is 60. But when n is larger, the precision will decrease, we think the reason may be the quality of web pages returned by search engine decreases. In addition, Figure 1 shows IARE-SV is better than SV 13% in the best precision result.

Figure 2 is the average value of recall of SV and IARE-SV in all data. Figure 2 shows that with the increase of n, the recall of classification get higher. When n is 40， IARE-SV get the max value of 82%, and SV get the max value of 78% when n is 60. IARE-SV raises 4% in the best recall result.

[1] www.qq.com

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[3] http://top.baidu.com/

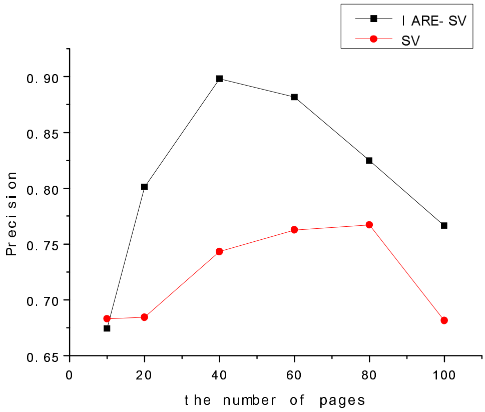


Fig. 1. The precise value of different number of pages

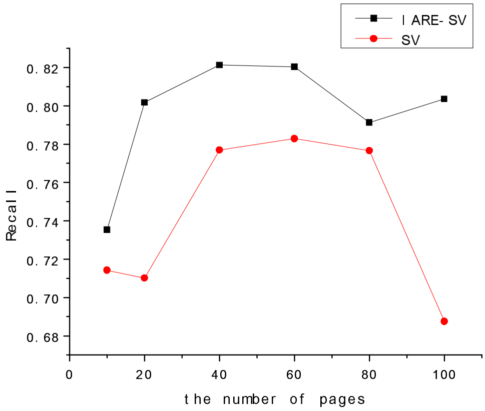


Fig. 2. The recall value of different number of pages

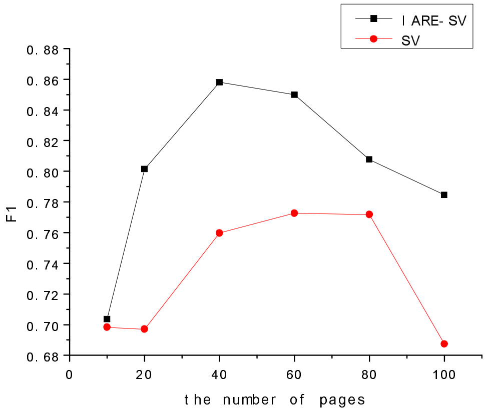


Fig. 3. The F1-measure value of different number of pages

Figure 3 is the result of F1-measure in recall and precision, IARE-SV is better than SV in all the value of *n*. When *n* is 40, IARE-SV get the max value of 84%, SV's max value is only 77% when *n* is 80. IARE-SV raises the result 7% in the best F1- measure result and need less enriched features. Thus, IARE-SV is more effective than SV.

6 Conclusion

We propose a query classification method based on index association rules expansion in this paper, according to the structure of search engine. Our method transforms online query classification to offline index term classification to improve the response time. In order to get better precision of index term enrichment, we propose novel algorithm which called index association expansion based on similarity voting (IARE-SV) to determine the category labels of index term. After feature words which have high support are selected to expand the index term, the similarity of the index term expansion and the web pages acts as the weight to select good web pages as feature enrichment of the index term. The experiment results show that our method can get better results in recall, precision and F1-measure than the simple voting method. However, because we have no actual search engine environment, our approaches can't be tested in large scale data. In future work, we will experiment with large scale data to evaluate our approaches.

References

1. Hu, J., Wang, G., Lochovsky, F., Sun, J.-t., Chen, Z.: Understanding user's query intent with wikipedia. In: Proceedings of the 18th International Conference on World Wide Web, pp. 471–480. ACM, Madrid (2009)
2. Li, X., Wang, Y.-Y., Shen, D., Acero, A.: Learning with click graph for query intent classification. ACM Trans. Inf. Syst. 28, 1–20 (2010)
3. Wang, C.-J., Lin, K.H.-Y., Chen, H.-H.: Intent boundary detection in search query logs. In: Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 749–750. ACM, Geneva (2010)
4. Shen, D., Li, Y., Li, X., Zhou, D.: Product query classification. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 741–750. ACM, Hong Kong (2009)
5. Broder, A.: A taxonomy of web search. SIGIR Forum 36, 3–10 (2002)
6. Shen, D., Pan, R., Sun, J.-T., Pan, J.J., Wu, K., Yin, J., Yang, Q.: Q2C@UST: our winning solution to query classification in KDDCUP 2005. SIGKDD Explor. Newsl. 7, 100–110 (2005)
7. Shen, D., Sun, J.-T., Yang, Q., Chen, Z.: Building bridges for web query classification. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 131–138. ACM, Seattle (2006)
8. Gabrilovich, E., Broder, A., Fontoura, M., Joshi, A., Josifovski, V., Riedel, L., Zhang, T.: Classifying search queries using the Web as a source of knowledge. ACM Trans. Web 3, 1–28 (2009)
9. Ganti, V., König, A.C., Li, X.: Precomputing search features for fast and accurate query classification. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 61–70. ACM, New York (2010)

10. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O.: Varying approaches to topical web query classification. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 783–784. ACM, Amsterdam (2007)
11. Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O.: Analysis of varying approaches to topical web query classification. In: Proceedings of the 3rd International Conference on Scalable Information Systems, pp. 1–5. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Vico Equense (2008)
12. Broder, A.Z., Fontoura, M., Gabrilovich, E., Joshi, A., Josifovski, V., Zhang, T.: Robust classification of rare queries using web knowledge. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 231–238. ACM, Amsterdam (2007)
13. Zhai, H., Guo, J., Wu, Q., Cheng, X., Sheng, H., Zhang, J.: Query Classification Based on Regularized Correlated Topic Model. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 552–555. IEEE Computer Society, Los Alamitos (2009)
14. Huang, M.X., Yang, X.W., Zhang, S.C.: Query Expansion of Pseudo Relevance Feedback Based on Matrix-Weighted Association Rule Mining. Journal of Software 20, 1854–1865 (2009)

A Link Analysis Model Based on Online Social Networks

Bu Zhan and ZhengYou Xia

Department of Computer, Nanjing University of Aeronautics and Astronautic, Nanjing, China
zhengyou_xia@nuaa.edu.cn

Abstract. As information technology has advanced, people are turning to electronic media more frequently for communication, and social relationships are increasingly found on online channels. Traditional on-line social network researches are based a certain comment interaction. Though some interest conclusions have been obtained, the understanding of the entire on-line social network is one-sided. In this paper, we compare four different types of networks proposed by previous researchers. Statistical analysis reveals that those four networks are consistent in nature (both the "small-world effect" and skewed degree distributions are found in them). To discover the mechanism behind these network observations, we propose a single-factor model with a single parameter κ; using this model, various networks can be obtained when we change the parameter κ in a given range. Simulation experiment based on this model show that the simulation results and the real data are consistent, which means that our model is valid.

1 Introduction

Although the first message boards (USENET) date back to 1979, the social networks defined by the comment interaction among their users have been studied only recently [1, 2]. Unlike personal weblogs, which receive few replies, the threads of BBS discussions reveal a rich complexity in their structure. Studies of USENET and other social-network services, such as Friendster and MySpace, have focused mainly on visualization techniques for understanding their social and semantic structures [2, 3]. Kou and Zhang proposed a model of interest space and showed that hierarchical and clustering structure in the interest space of BBS reply networks exhibits not only small-world characteristics but also preferential attachment, which is a common explanation for their scale-free properties [4]. K.I. Goh and co-workers proposed a simple network model to identify the chief statistical characteristics of message-board networks [1]. In Vicenc Gomez' work, a simple measure is used to evaluate the degree of controversy provoked by a post [2]. In previous studies of on-line social network, the research process follows the old stereotype: proposing some kind of network construction method; taking necessary statistical analysis; summarizing different statistics to get corresponding conclusions. Sometimes, community structure and network evolution are incidentally considered. Those network construction methods are basically based on two types: comment interaction and comment content. The former has been studied extensively [1, 2, 4], while the latter attracts increasing attentions recently [5, 6]. Though some interest conclusions have been obtained, the understanding of the entire on-line social network is one-sided. Here, we compare

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 319–326, 2011.

four different types of networks proposed by previous researchers. Statistical analysis reveals that no matter what kind of construction method we take, those four networks are consistent in nature (both the “small-world effect” and skewed degree distributions are found in them). Only subtle variations exist in clustering function, nearest-neighbor degree distribution, strength distribution, etc. To confirm this conclusion, we propose a single-factor model with a single parameter κ; using this model, various networks can be obtained when we change the parameter κ in a given range. Simulation experiment based on this model show that the simulation results are consistent the real data, which means that our model is valid.

2 Networks Derived from BBS

To study the Tianya social network, we began by adopting the formalism in [2]. Every registered user identification (ID) corresponds to a node i∈V in a graph G=<V, E>. An edge (i, j)∈E represents a social relation between two users that results from their comment activity. Let n_{ij} be the number of times that user i writes a comment to user j. Links between two users can be defined using n_{ij} in several ways. In Vicenc Gomez’ work, three network types were discussed systematically; here, we focus on undirected networks defined as:

Undirected dense network: An undirected edge exists between users i and j if either n_{ij}>0 or n_{ji}>0. The weight of that edge w_{ij} equals the sum n_{ij}+n_{ji}.

Undirected sparse network: An undirected edge exists between users i and j if n_{ij}>0 and n_{ji}>0. The weight of that edge w_{ij} is defined as w_{ij}=min(n_{ij}, n_{ji}).

We propose two additional networks, one based on the comment interaction, the other based on the comment content:

Interest network: An undirected edge exists between users i and j if if n_{ip}>0 and if n_{jp}>0 where user p began a thread. The weight of an edge w_{ij} is defined as $w_{ij} = \sum_{p\in P} \min(n_{ip}, n_{jp})$ where P is the set of users who have begun a thread. The average number of follow-up postings on Tianya is about 50 per thread; thus, the total number of network edges is $O(n_P \cdot 50^2)$ where n_P is the number of BBS threads. Since that number is huge, we introduce a threshold k, such that an undirected edge exists if and only if w_{ij}>k.

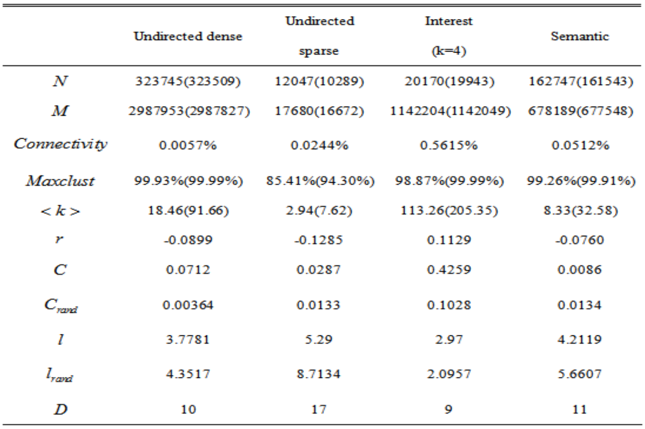
Semantic network: An undirected edge exists between users i and j if n_{ij}>0 and n_{ji}>0 and also if there are emotional words in the replies between them. Emotional words include two basic types: supportive and opposing. For example, phrases such as “好贴，顶起来”, “支持楼主” or “经典” are supportive, whereas words such as “NND”, “TM” or “鄙视” are opposing. We identified roughly 50 terms or phrases from the public discussions on Tianya.com as either supportive or opposing and assigned each a value between 0 and 1 according to their tone. A higher value corresponds to a greater degree of support; every phrase has an associated numerical “trust”. Thus, the semantic network is defined as follows:

1) The undirected dense network is initialized with every edge given a “trust” value of 0.5;
2) Analyzing every comment between two users and updating the “trust” value of every edge (if there are several emotional words in one comment, we can roughly take average);
3) Discarding edges whose “trust” values between $\boxed{0.5-\xi}$ and $\boxed{0.5+\xi}$ ($\boxed{\xi}$ will be seen as a threshold).

3 Statistical Analyses of the Four BBS Networks

Here, we discuss network characteristics from a global perspective. Table 1 shows the statistics describing the four networks.

Table 1. Statistics of the Tianya social networks



	Undirected dense	Undirected sparse	Interest (k=4)	Semantic
N	323745(323509)	12047(10289)	20170(19943)	162747(161543)
M	2987953(2987827)	17680(16672)	1142204(1142049)	678189(677548)
Connectivity	0.0057%	0.0244%	0.5615%	0.0512%
Maxclust	99.93%(99.99%)	85.41%(94.30%)	98.87%(99.99%)	99.26%(99.91%)
$<k>$	18.46(91.66)	2.94(7.62)	113.26(205.35)	8.33(32.58)
r	-0.0899	-0.1285	0.1129	-0.0760
C	0.0712	0.0287	0.4259	0.0086
C_{rand}	0.00364	0.0133	0.1028	0.0134
l	3.7781	5.29	2.97	4.2119
l_{rand}	4.3517	8.7134	2.0957	5.6607
D	10	17	9	11

We see that, for all graphs, The average clustering coefficient C is much higher than the randomized counterpart Crand [9,10]. Large C values indicate that discussions can be begun among bunches of users easily. Small l values indicate that ideas and opinions can propagate rapidly from one person to another. Hence, the small-world topologies of BBS networks ensure the propagation of discussions among users. This is consistent with other real-world networks, which exhibit similar deviations from the random graph, and adds to the collection of networks having the “small-world” property [11]. Table 1 also shows the correlation coefficient r [12,13] (also called the Pearson correlation coefficient) for our three networks. Interestingly, unlike traditional social networks, which exhibit significant assortative mixing, the undirected dense network and the undirected sparse network are characterized by disassortative mixing. In the interest network, the Pearson correlation coefficient is greater than 0.1, meaning that the interest network is assortative

The most basic topological characterization of a graph G can be expressed in terms of its degree distribution P(k), which is defined as the probability that a node chosen uniformly at random has degree k or, equivalently, as the fraction of nodes in the graph having degree k. The analysis of this degree distribution describes the level of interaction between users and provides a robust indicator of the degree of heterogeneity within the network. In this section, we discuss the degree distributions of the four Tianya networks.

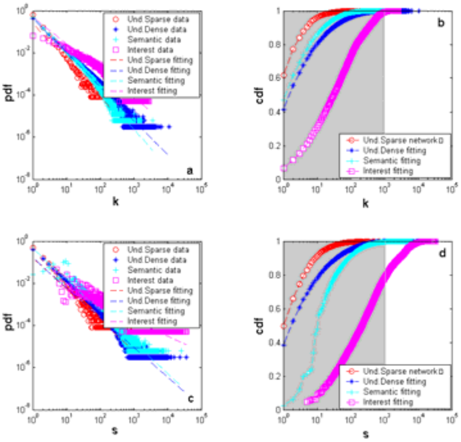


Fig. 1. Degree and strength distributions of the four networks

The degree k_i of a user i is distributed according to a power law followed by an exponential cutoff , as shown in Fig. 1(a). The cumulative distribution function (cdf) of the degrees is shown in Fig. 1(b). The strength s_i of a user i is the sum of the weight of each edge attached to i . As shown in Fig. 1(c), the strength distribution is also followed by an exponential cutoff.

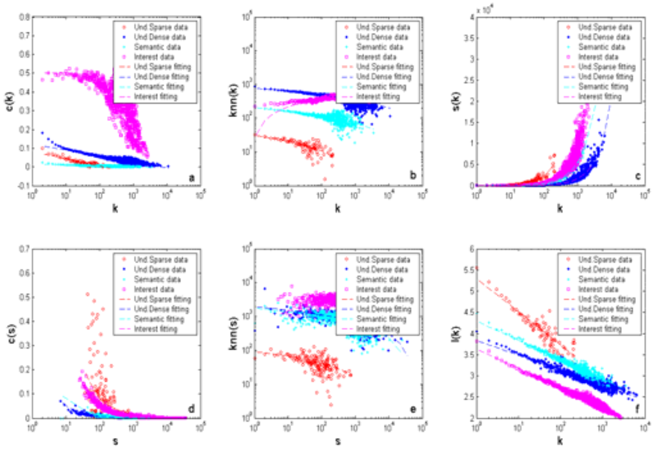


Fig. 2. c(k), knn(k), c(s), knn(s) of the four networks

The clustering function C(k) is defined as the average of C_i over all vertices with a given degree k. For the undirected dense network, the undirected sparse network and the semantic network, C(k) decays as $\alpha \log(k) + \beta$, with $\alpha < 0$, which is consistent with Ref. [10]. However, for the interest network, C(k) decays as $\alpha \cdot e^{\beta \cdot k}$, with $\beta < 0$. The average nearest-neighbor degree function $knn(k)$ [14,15], which is defined as the average degree of the neighbors of vertices of degree k, also follows a logarithmic distribution for all of these networks. The upward curvature of knn(k) obeys the Matthew effect, implying that the friends of the hub members are more likely to be hub members themselves, in agreement with typical social networks. The nonlinear relationship between s and k implies that hub members tend to post messages considerably more frequently than other people. The average shortest-path degree function l(k) obeys a logarithmic distribution, which means that hub members are more likely to be acquainted with other people.

4 A New Model

After analyzing the chief statistical characteristics of the four networks, we reach the qualitative conclusion that those four networks are consistent in nature (both the “small-world effect” and skewed degree distributions are found in them). Only subtle variations exist in clustering function, nearest-neighbor degree distribution, strength distribution, etc. To test these results and to help form a more complete picture of the properties of BBS networks, we carried out computer simulations on a simple network model. First, we randomly generated N nodes and assigned every node a degree according to a desired degree distribution. Consider the symmetric binomial form $e_{ij} = \sigma(\frac{k_i \cdot k_j}{k_{\max}^2})^{\kappa}$, where e_{ij} represents the probability that an edge exists between a node of degree i and a node of degree j, with $\kappa > 0$, and where

$$\sigma = \frac{\sum_i k_i \cdot P(k_i)}{(N-1) \cdot \sum_{i<j} (\frac{k_i \cdot k_j}{k_{\max}^2})^{\kappa}}$$

is a normalization constant. By changing the value of κ, we generated different network topologies and studied their corresponding characteristics such as <k>, C, P and L. To validate our model, we compared these simulated results with the real data of our database. Figure 3 shows the topology changes of different network types as κ is varied from 0 to 10; here, N equals 100000. As shown in Fig. 3(a), we obtain the best fit for the undirected sparse network when κ lies between 0.1 and 0.2. The best-fitting interval for the undirected dense and semantic networks lies between 0.1 and 1, as shown in Fig. 3(b, c). By contrast, the best-fitting κ for the interest network is significantly different from the other three values, being greater than 1 as shown in Fig. 3(d).

Table 2. The maximum values for the four network types

	Undirected dense	Undirected sparse	Interest (k=4)	Semantic
Maxavgdegree	56.472	4.7292	137.35	57.347
Maxcluster	0.49967	0.44967	0.57489	0.31866
Maxabs(pearson)	-0.053548	-0.10355	0.14819	-0.033176
Maxavgshortpath	3.6118	5.4118	5.1323	3.9646

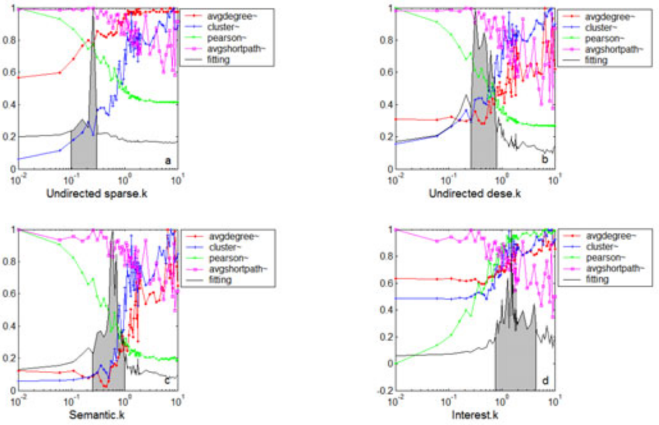


Fig. 3. Plot of various statistics as K is varied for the four networks. Here, avgdegree ~ represents the ratio of the current avgdegree of value K to the maximal avgdegree; the rest may be deduced by analogy. The maximum values for the four networks are given in Table 3.

To verify the validity of our model, we conducted a large number of computer simulations and property comparisons: we generate four artificial networks with the model; then statistical analysis is implemented; property comparison is carried out between the real data and simulation data finally.

Figure 4 shows comparison results between the real data and the simulation data for the four networks respectively. As we discussed in Section 3, The clustering function C(k) quantifies how well connected are the neighbors of a vertex in a graph, which has been taken as a signature of the network hierarchical structure. And the average nearest-neighbor degree function knn(k) is a vital distribution which can lead to the Pearson correlation coefficients. The distributions of these two functions are almost overlap: they all obeys the logarithmic distribution, and parameters in the logarithmic formula change little between the real data and the simulative one, which means that our model can restore them beautifully. While for the average shortest-path degree function l(k), our model performes just passable: although the relative trends of the real data and the simulation data stay the same, the later is always better than the former, which means the artificial networks with our model are over idealistic, some random factors in real networks may not be considered. This will be further improved later.

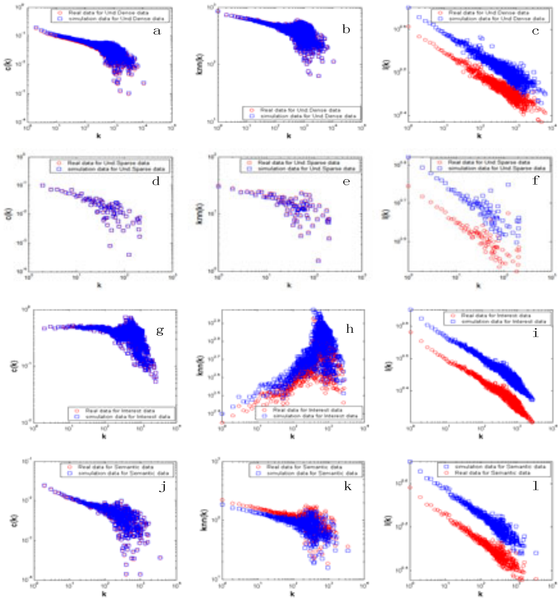


Fig. 4. The Comparison between real data and simulation data in the clustering function, the nearest-neighbor degree distribution and the strength distribution. (a, b, c) Comparison results of the undirected dense network. (d, e, f) Comparison results of the undirected sparse network. (g, h, i) Comparison results of the interest network. (j, k, l) Comparison results of the semantic network.

5 Conclusions and Discussion

Previous researches are basically based on two types: comment interaction and comment content. In this paper, we prove that no matter what kind of construction method we take, networks constructed according these two type are consistent in nature: both the “small-world effect” and skewed degree distributions are found in them. Only subtle variations exist in clustering function, nearest-neighbor degree distribution, strength distribution, etc. To test these results, we proposed a simple network model. For different networks, P(k) is fixed while the parameter κ is changed to obtain different network topologies. As seen in Fig. 3, when κ is lower than 0.2, our model resembles the undirected sparse network; if κ lies between 0.1 and 1, it resembles the undirected dense network or the semantic network; and when κ is greater than 1, the model resembles the interest network. Simulation experiment based on this model show that the simulation results and the real data are consistent largely, which means that our model is valid. The model not only helps to characterize different network types, but can also provide a more complete picture of the properties of BBS networks.

Acknowledgments. Our work is supported by the ministry of science and technology of the People's Republic of China NO 2009GJE00035 and NUAA foundation founds NS2010111.

References

[1] Goh, K.-I., Eom, Y.-H., Jeong, H., Kahng, B., Kim, D.: Structure and evolution of online social relationships: Heterogeneity in unrestricted discussions. Phys. Rev. E 73, 066123 (2006)

[2] Gomez, V., Kaltenbrunner, A., Lopez, V.: Statistical analysis of the social network and discussion threads in slashdot. In: WWW, Beijing, China, April 21-25 (2008)

[3] Chen, I.-X., Yang, C.-Z.: Visualization of Social Networks. In: Handbook of Social Network Technologies and Applications, Part 5, pp. 585–610 (2010)

[4] Kou, Z., Zhang, C.: Reply networks on a bulletin board system. Phys. Rev. E 67, 036117 (2003)

[5] Mishne, G., Glance, N.: Leave a reply: an analysis of weblog comments. In: WWW, May 22-26 (2006)

[6] Yano, T., Smith, N.A.: What's worthy of comment? Content and comment volume in political blogs. In: 4th Int'l AAAI Conference on Weblogs and Social Media (2010)

[7] Watts, D.J.: Small Worlds. Princeton University Press, Princeton (1999)

[8] Newman, M.E.J.: Assortative mixing in networks. Phys. Rev. Lett. 89, 208701 (2087)

[9] Bo, R., Xia, Z., Yongzhen, Z., Bu, Z.: Research on BBS Complex Online Network and Members Interactive Characteristics (2009) (Chinese version)

[10] Soffer, S.N., Vazquez, A.: Network clustering coefficient without degree-correlation biases. Phys. Rev. E 71, 051701 (2005)

[11] Goldberg, M., Kelley, S., Magdon-Ismail, M., Mertsalov, K., Wallace, W.A.: Communication Dynamics of Blog Networks. In: Giles, L., Smith, M., Yen, J., Zhang, H. (eds.) SNAKDD 2008. LNCS, vol. 5498, pp. 36–54. Springer, Heidelberg (2010)

[12] Newman, M.E.J.: Mixing patterns in networks. Phys. Rev. E 67, 026126 (2003)

[13] Rangwala, H., Jamali, S.: Defining a Coparticipation Network Using Comments on Digg. In: Association for the Advancement of Artificial Intelligence (2010)

[14] Boguna, M., Pastor-Satorras, R.: Epidemic spreading in correlated complex networks. Phys. Rev. E 66, 047104 (2002)

[15] Boguna, M., Pastor-Satorras, R., Vespignani, A.: Absence of epidemic threshold in scale-free networks with degree correlations. Lect. Notes Phys. 625, 127 (2003)

[16] Xia, Z.Y.: Fighting criminals: Adaptive inferring and choosing the next investigative objects in the criminal network. Knowledge-Based Systems (2008)

[17] Xia, Z., Wang, J.: DIMH: A novel model to detect and isolate malicious hosts for mobile ad hoc network. Computer Standards & Interfaces, 660–669 (2006)

[18] Xia, Z.: A Novel Policy and Information Flow Security Model for Active Network. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.) ISI 2004. LNCS, vol. 3073, pp. 42–55. Springer, Heidelberg (2004)

[19] Luo, X., Xu, Z., Yu, J., Chen, X.: Building Association Link Network for Semantic Link on WebResources. IEEE Transactions on Automation Scienceand Engineering (forthcoming)

Research on Information Measurement at Semantic Level

Kaizhong Jiang, Lu Li, and Bosheng Xu

College of Basic Teaching
Shanghai University of Engineering Science
Shanghai, China
{kzjiang,lilu,xbs0107}@sues.edu.cn

Abstract. The paper defined an information measure associated with a topic or semantics for a keyword based corpus. Firstly, the topic-based corpus was obtained. Then the latent semantic vector space model of the corpus was established. After that, the information measure of the keyword was defined through the vector space model. Accordingly, it could be calculated that the amount of the topic information any document contained. Lastly, the membership degree which measured the degree of membership of the document belonging to the topic was introduced. Set a measurement threshold, thereby it was determined whether the documents belonging to the topic or not. Experiments show that the definition of the information measurement can get over the difficulty of the word-match search and real reach the goal of the Semantic-match search.

Keywords: Latent Semantics, information measurement, metric distribution, membership degree.

1 Introduction

Petri pointed out that the information should be regarded as physical quantity like the energy, momentum, angular momentum, and should have its own conservation law and other laws of physics. Petri considered that be a new discipline, referred as the information dynamics [1].

Well-known shannon's information theory is based on the statistical theory and take the information entropy as the amount of information. It is triumphantly used in some fields, such as telecommunication technology [2]. However, a lot of scholars, even Shannon himself, realized shannon's theory can not meet the needs of many fields [3].

W. Weaver pointed out that there are three levels of communication: the technical level, the semantic level and the utility level of communication [4]. Shannon's information theory is at technical level.

In order to consider the semantics of information, R. Carnap et al put forward a method which used logical probability with the replacement of ordinary (random) probability and then measured semantic information with shannon's entropy of information content [5]. However, this method has many restrictions required and therefore of little value.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 327–334, 2011.

Considering the fuzziness of information, fuzzy sets, rough sets and other sets were presented [6-7].

Currently, the most information retrieval systems are based on the word-match method [8]. However, because of the broadly synonymy and polysemy problem, the returned documents may not match with the conceptual topic that user want. All the semantic-based search algorithm must solve this problem [8]. Therefore, it is necessary to study on the semantic information measure of the keywords.

The so-called semantic-based retrieval is defined as the retrieval that the retrieved information and the user's query must have the same conceptual topic, but may not have the common terms or keywords [9].

T. K. Landauer [10] et al take shannon’s measure method of the information as the measure method of the semantic information of keywords based on corpus.

Shannon's theory can not be copied to the knowledge of the macro-information structure, because Shannon's study is based on the bit level [1]. The information theory over bit level, such as semantic level, should require more research.

The paper studies on the semantic information measure of the keyword of the corpus. The corpus is based on a topic. So, there should be a significant association between the keyword’s information measure and the topic. The information measure like this is the semantic information measure. The same keyword has different semantic information measures in different topic corpora.

The organization of the paper is as follows. In Section 2, briefly review the vector space modelling of Latent Semantic Indexing (LSI) model of information retrieval. The semantic information measure is defined and analyzed in section 3. In Section 4, the membership degree of any document belonging to theme C is defined. Test and evaluation are in the section 5. Finally, we conclude the paper in Section 6.

If no special instructions, corpus, corpus theme, document-keyword matrix are all expressed as **C**. Bold letters denote vectors or matrices.

2 Vector Space Modelling

LSI is an important extension of the Vector Space Model (VSM) and the one of most efficient information retrieval models in recent years [11-12]. 1990, Deerwester et al applied the singular value decomposition (SVD) to information retrieval [8]. Then LSI become a new research hotspot [13-14]. LSI takes advantage of the implicit relationship of terms within a document to improve the retrieval performance. Furthermore, LSI is a method of dimensionality reduction [15].

Let C={$\mathrm{d}_1,\cdots,\mathrm{d}_m$} be the corpus with respect to topic C, T={$t_1,\cdots,t_L$} be the set of all the keywords of the corpus C, $f_i(t_j)$ be the frequency with which keyword t_j occurs in the d_i, $f(t_j)$ be the total number of the documents including t_j in the corpus.

The i-th document d_i is expressed as a vector: $\mathbf{d}_i=(w_{i1},\ldots,w_{iL})$, where w_{ij} are the TFIDF weight values, i.e., $w_{ij}=l(i,j)\cdot g(j)$.

$l(i,j)$, called the local weight, denotes the ability strength of t_j to describe the whole document d_i. Generally, take $f_i(t_j)$ as $l(i,j)$, i.e., $l(i,j)= f_i(t_j)$. $g(j)$, called the global weight, denotes the ability strength of t_j to distinguish between the documents. Take the Inverse Document Frequency (IDF) of t_j as $g(j)$, i.e., $g(j)=\log_2 m-\log_2 f(t_j)$.

The j-th keyword t_j is expressed as a vector: $\mathbf{T}_j=(w_{1j},\ldots,w_{mj})^{\mathrm{T}}$ correspondingly. $\mathbf{T}_j$ reflect the weight distribution of the t_j in the corpus.

Definition 1 Core Keyword. Given a threshold value $w_{\min}$, if $\sum_{i=1}^{m} w_{ij} > w_{min}$, t_j is called a corpus core keyword, core keyword for short.

Might as well supposition $\mathrm{T_C}=\{t_1,\cdots, t_n\}(n<<L)$ is the core keyword set.

Let $\mathbf{C}=(w_{ij})_{m\times n}=(\mathbf{d}_1,\cdots, \mathbf{d}_m)^{\mathrm{T}}$, called the keyword-document matrix of the corpus C.

In order to make documents of different lengths equal importance in the same corpus, document vector must be normalized as follows:

$\|\mathbf{d}_i\|=\sqrt{w_{i1}^2+\cdots+w_{in}^2}$, $\mathbf{d}_i^0=\mathbf{d}_i/\|\mathbf{d}_i\|=(w_{i1}{}^0,\ldots,w_{in}{}^0)$, $\mathrm{C}^0=(w_{in}{}^0)_{m\times n}$.

For convenience, the following document vector $\mathbf{d}_i$ and keyword-document matrix **C** all denote the normalized vectors and matrix. Keyword denote the core keyword.

Definition 2 Partial order relation. Let t_j and t_k be any two core keywords, $\mathbf{T}_j=(w_{1j},\ldots,w_{mj})^{\mathrm{T}}$ and $\mathbf{T}_k=(w_{1k},\ldots,w_{mk})^{\mathrm{T}}$ be two vectors of t_j and t_k respectively. $J=\{i|w_{ij}>0, i=1,\ldots,m\}$, $K=\{i|w_{ik}>0, i=1,\ldots,m\}$,

(1) t_k is said to include t_j, (denoted by $t_j\subseteq t_k$), if and only if $J\subseteq K$.

(2) $\mathbf{T}_j$ is said to be less than or equal to $\mathbf{T}_k$, (denoted by $\mathbf{T}_j\preccurlyeq\mathbf{T}_k$), if and only if $w_{ij}\le w_{ik}$ for any $i(1\le i\le m)$.

Obviously, $\subseteq$ is a partial order relation of core keyword set $\mathrm{T_C}$, and $\preccurlyeq$ is a partial order relation of core keyword vector set $\mathbf{T}_{\mathrm{C}}$.

For any two keywords t_j and t_k, the following triangle inequality and Cauchy inequality are always true, i.e.

$\|\mathbf{T}_j+\mathbf{T}_k\|\le\|\mathbf{T}_j\|+\|\mathbf{T}_k\|$; $0\le\mathbf{T}_j\cdot\mathbf{T}_k\le\|\mathbf{T}_j\|\cdot\|\mathbf{T}_k\|$. where $\mathbf{T}_j\cdot\mathbf{T}_k$ denotes the dot product.

Definition 3 Latent Semantic Similarity Degree. For any two keywords t_j and t_k, $S(t_j,t_k)=\mathbf{T}_j\cdot\mathbf{T}_k/(\|\mathbf{T}_j\|\cdot\|\mathbf{T}_k\|)$, call $S(t_j,t_k)$ the latent semantic similarity degree of t_j and t_k.

t_j and t_k are said to be approximately latent semantic similarity, if there exists a small threshold value $\varepsilon\ge0$ such that $\|\mathbf{T}_j\|\cdot\|\mathbf{T}_k\|-\mathbf{T}_j\cdot\mathbf{T}_k\le\varepsilon$.

If t_j and t_k are approximately latent semantic similarity, there must exist a real number $a>0$, such that $\mathbf{T}_j\approx a\mathbf{T}_k$, and vice versa.

3 Semantic Information Measure

For a document, it is the use of the corpus core words that make the document's theme converge to the corpus theme.

Definition 4 Semantic Information Measurement. Let $\mathbf{C}=(w_{ij})_{m\times n}$ denote the row normalized keyword-document matrix, $\mathbf{T}_j=(w_{1j},\ldots,w_{mj})^{\mathrm{T}}$ the vector of the t_j correspondingly, $\mathrm{I}(t_j)=\|\mathbf{T}_j\|=\sqrt{w_{1j}^2+\cdots+w_{mj}^2}$, call $\mathrm{I}(t_j)$ the semantic information measurement of keyword t_j.

Because it is unlike other similar literatures that $\mathrm{I}(t_j)$ is taken as the semantic information measurement of keyword and $\mathrm{I}(t_j)$ runs through the paper, the following discussion of $\mathrm{I}(t_j)$ will be necessary.

Let $\mathbf{T}_j = (w_{1j},\ldots,w_{mj})^{\mathrm{T}}$ be the keyword vector of the t_j, where

$$w_{ij} = f_i(t_j)\log_2\frac{m}{f(t_j)} \bigg/ \sqrt{\sum\nolimits_{k=1}^{n}\left(f_i(t_k)\log_2\frac{m}{f(t_k)}\right)^2} \quad (\,i=1,\cdots,m).$$

$\mathbf{T}_j$ reflects the weight distribution of t_j on the documents in the corpus C, as illustrated in Figure.1(might as well supposition the sort order is descending by t_j's weight).

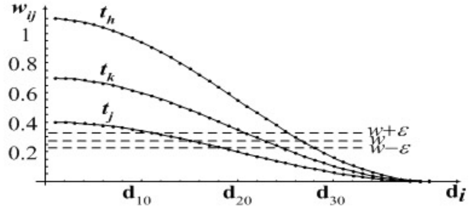


Fig. 1. Weight distribution of keyword on the documents

For any two keywords t_j and t_k, let
$J=\{i|w_{ij}>0,\ i=1,\ldots,m\}$, $K=\{i|w_{ik}>0,\ i=1,\ldots,m\}$, $|J-K|=n_1$, $|K-J|=n_2$, $|J\cap K|=n_3$, $\mathbf{e}_i=(0,\ldots,1,\ldots,0)^{\mathrm{T}}$, where only i-th component is 1, $V^{(j-k)}=span\{\ \mathbf{e}_i \mid i\in J-K\}$, $V^{(k-j)}=span\{\ \mathbf{e}_i \mid i\in K-J\}$, $V^{(j\cap k)}=span\{\ \mathbf{e}_i \mid i\in K\cap J\}$, then $\mathbf{T}_j$ can be decomposed two parts: $\mathbf{T}_j=(w_{1j},\ldots,w_{mj})^{\mathrm{T}}=\ \mathbf{T}_j^{(j-k)}+\mathbf{T}_j^{(j\cap k)}$, where $\mathbf{T}_j^{(j-k)}=\sum_{i\in J-K} w_{ij}\mathbf{e}_i$, $\mathbf{T}_j^{(j\cap k)}=\sum_{i\in K\cap J} w_{ij}\mathbf{e}_i$ $(i\in J)$. As a matte of fact, $\mathbf{T}_j^{(j-k)}$ and $\mathbf{T}_j^{(j\cap k)}$are the projection of the $\mathbf{T}_j$ on space $V^{(j-k)}$ and $V^{(j\cap k)}$ respectively.

Similarly, $\mathbf{T}_k = \mathbf{T}_k^{(k-j)} + \mathbf{T}_k^{(j\cap k)}$, where $\mathbf{T}_k^{(k-j)}=\sum_{i\in K-J} w_{ik}\mathbf{e}_i$, $\mathbf{T}_k^{(j\cap k)} = \sum_{i\in K\cap J} w_{ik}\mathbf{e}_i$ $(i\in K)$, $\mathbf{T}_k^{(k-j)}$ and $\mathbf{T}_k^{(j\cap k)}$ are the projection of the $\mathbf{T}_k$ on space $V^{(k-j)}$ and $V^{(j\cap k)}$.

Definition 5 Add operation. For any two keywords t_j and t_k, J and K as definition 2, n_1, n_2, n_3 as above, if t_j and t_k are to be taken as one abstract word, represented as $t_h= t_j\uplus t_k$, its vector denoted as $\mathbf{T}_h= (w_{1h},\ldots,w_{mh})^{\mathrm{T}}$, where

$$w_{ih} = \frac{f_i(t_h)\log_2\dfrac{m}{f(t_h)}}{\sqrt{\sum_{l\neq j,k}\left(f_i(t_l)\log_2\dfrac{m}{f(t_l)}\right)^2+\left(f_i(t_h)\log_2\dfrac{m}{f(t_h)}\right)^2}}\ ,\quad f_i(t_h)=\begin{cases} f_i(t_j) & i\in J-K \\ f_i(t_j)+f_i(t_k) & i\in J\cap K \\ f_i(t_k) & i\in K-J \\ 0 & i\notin J\cup K \end{cases} \quad \text{and}$$

$f(t_h)= n_1+n_2+n_3, f(t_j)= n_1+ n_3, f(t_k)= n_2+n_3$.

Corollary 1. For any two keywords t_j and t_k, $t_h= t_j\uplus t_k$, $\mathbf{T}_j=(w_{1j},\ldots,w_{mj})^{\mathrm{T}}$, $\mathbf{T}_k=(w_{1k},\ldots,w_{mk})^{\mathrm{T}}$ and $\mathbf{T}_h=(w_{1h},\ldots,w_{mh})^{\mathrm{T}}$ are their vectors respectively, then $\|\mathbf{T}_h\|\le\|\mathbf{T}_j\|+\|\mathbf{T}_k\|$.

Proof: $\mathbf{T}_h$ can be decomposed three parts: $\mathbf{T}_h = \sum_{i\in J\cup K} w_{ih}\mathbf{e}_i = \mathbf{T}_h^{(j-k)} + \mathbf{T}_h^{(k-j)} + \mathbf{T}_h^{(j\cap k)}$, where $\mathbf{T}_h^{(j-k)}= \sum_{i\in J-K} w_{ih}\mathbf{e}_i$, $\mathbf{T}_h^{(k-j)}= \sum_{i\in K-J} w_{ih}\mathbf{e}_i$, $\mathbf{T}_h^{(j\cap k)}= \sum_{i\in K\cap J} w_{ih}\mathbf{e}_i$.

When $i\in J\text{-}K, f_i(t_k)$=0, so w_{ik}=0. $f(t_h)\geq f(t_j)$, $f_i(t_h)\ln\frac{m}{f(t_h)}\leq f_i(t_j)\ln\frac{m}{f(t_j)}$.

The function $y=\frac{x}{\sqrt{c+x^2}}$ (c>0, x>0) is a positive and strictly increasing function,

so $$w_{ih}=\frac{f_i(t_h)\log_2\frac{m}{f(t_h)}}{\sqrt{\sum\limits_{l\neq j,k}\left(f_i(t_l)\log_2\frac{m}{f(t_l)}\right)^2+\left(f_i(t_h)\log_2\frac{m}{f(t_h)}\right)^2}}\leq\frac{f_i(t_j)\log_2\frac{m}{f(t_j)}}{\sqrt{\sum\limits_{l\neq j,k}\left(f_i(t_l)\log_2\frac{m}{f(t_l)}\right)^2+\left(f_i(t_j)\log_2\frac{m}{f(t_j)}\right)^2}}=w_{ij}.$$

That is every component of $\mathbf{T}_h^{(j\text{-}k)}$ is less than or equal to corresponding component of $\mathbf{T}_j^{(j\text{-}k)}$.

Therefore, $\mathbf{T}_h^{(j\text{-}k)}\preccurlyeq\mathbf{T}_j^{(j\text{-}k)}$, consquently, $\|\mathbf{T}_h^{(j\text{-}k)}\|\leq\|\mathbf{T}_j^{(j\text{-}k)}\|$.

Similarly, when $i\in K\text{-}J$, $w_{ih}\leq w_{ik}$, $\mathbf{T}_h^{(k\text{-}j)}\preccurlyeq\mathbf{T}_k^{(k\text{-}j)}$, $\|\mathbf{T}_h^{(k\text{-}j)}\|\leq\|\mathbf{T}_k^{(k\text{-}j)}\|$.

when $i\in K\cap J, f_i(t_h)=f_i(t_j)+f_i(t_k)$, $f(t_h)\leq f(t_j)$, $f(t_h)\leq f(t_k)$, so,

$$f_i(t_h)\log_2\frac{m}{f(t_h)}\leq f_i(t_j)\log_2\frac{m}{f(t_j)}+f_i(t_k)\log_2\frac{m}{f(t_k)}.$$

$$w_{ih}=\frac{f_i(t_h)\log_2\frac{m}{f(t_h)}}{\sqrt{\sum\limits_{l\neq j,k}\left(f_i(t_l)\log_2\frac{m}{f(t_l)}\right)^2+\left(f_i(t_h)\log_2\frac{m}{f(t_h)}\right)^2}}\leq\frac{f_i(t_j)\log_2\frac{m}{f(t_j)}+f_i(t_k)\log_2\frac{m}{f(t_k)}}{\sqrt{\sum\limits_{l\neq j,k}\left(f_i(t_l)\log_2\frac{m}{f(t_l)}\right)^2+\left(f_i(t_j)\log_2\frac{m}{f(t_j)}+f_i(t_k)\log_2\frac{m}{f(t_k)}\right)^2}}$$

$$\leq\frac{f_i(t_j)\log_2\frac{m}{f(t_j)}+f_i(t_k)\log_2\frac{m}{f(t_k)}}{\sqrt{\sum\limits_{l\neq j,k}\left(f_i(t_l)\log_2\frac{m}{f(t_l)}\right)^2+\left(f_i(t_j)\log_2\frac{m}{f(t_j)}\right)^2+\left(f_i(t_k)\log_2\frac{m}{f(t_k)}\right)^2}}=w_{ij}+w_{ik}.$$

That is $\mathbf{T}_h^{(j\cap k)}\preccurlyeq\mathbf{T}_j^{(j\cap k)}+\mathbf{T}_k^{(j\cap k)}$. Therefore, $\|\mathbf{T}_h^{(j\cap k)}\|\leq\|\mathbf{T}_j^{(j\cap k)}+\mathbf{T}_k^{(j\cap k)}\|$.

As a whole, for any $i\in J\cup K$, $w_{ih}\leq w_{ij}+w_{ik}$, thus, $\mathbf{T}_h\preccurlyeq\mathbf{T}_j+\mathbf{T}_k$, $\|\mathbf{T}_h\|\leq\|\mathbf{T}_j+\mathbf{T}_k\|\leq\|\mathbf{T}_j\|+\|\mathbf{T}_k\|$.

Corollary 2. For any two keywords t_j and t_k, t_h= $t_j\uplus t_k$, $\mathbf{T}_j=(w_{1j},\ldots,w_{mj})^{\mathrm{T}}$, $\mathbf{T}_k=(w_{1k},\ldots,w_{mk})^{\mathrm{T}}$ and $\mathbf{T}_h=(w_{1h},\ldots,w_{mh})^{\mathrm{T}}$ are their vectors respectively, (1) If $t_k\subseteq t_j$, then $\|\mathbf{T}_j\|\leq\|\mathbf{T}_h\|$. (2) If $t_j\subseteq t_k$, then $\|\mathbf{T}_k\|\leq\|\mathbf{T}_h\|$. (3) If $t_k\subseteq t_j$ and $t_j\subseteq t_k$, then $\max\{\|\mathbf{T}_j\|,\|\mathbf{T}_k\|\}\leq\|\mathbf{T}_h\|$.

Proof is omitted here for limited space.

Case (3) as shown in Figure 1.

The results above mean that a keyword combined with its included keyword can increase the semantic information measurement.

Given a keyword t_j, a number w>0 and a small positive number ε>0, $L(t_j)$= $\{\mathbf{d}_i|w_{ij}>w+\varepsilon\}$, $U(t_j)=\{\mathbf{d}_i|w_{ij}>w-\varepsilon\}$, $B(t_j)=U(t_j)-L(t_j)=\{\mathbf{d}_i|w-\varepsilon<w_{ij}\leq w+\varepsilon\}$, as shown in Figure 1. Call $L(t_j)$ the keyword t_j's lower approximation $U(t_j)$ the t_j's upper approximation and $B(t_j)$ the t_j's border. The documents of the set $L(t_j)$ means the required documents searching with the keyword t_j and the documents of the set C−$U(t_j)$ means not. The smaller the $|B(t_j)|$ (cardinal number) is, the stronger the ability of t_j's distinguishing documents is, and vice versa.

It is assumed that t_j and t_k are latent semantic similar based on C, as shown in Fig 1, i.e., $\mathbf{T}_k= (w_{1k},\ldots,w_{mk})^{\mathrm{T}}\approx (w_{1j},\ldots,w_{mj})^{\mathrm{T}} =a\mathbf{T}_j$ ($a>0$). Therefore, $\mathrm{I}(t_k)\approx a\mathrm{I}(t_j)$.

It is assumed that t_j and t_k are latent semantic similar based on C, i.e., $\mathbf{T}_k\approx a\mathbf{T}_j$. Take $t_j\uplus t_k$ as a abstract keyword t_h, whose keyword vector $\mathbf{T}_h$ is larger than or equal to $\mathbf{T}_j$ or $\mathbf{T}_k$. So, $t_j\uplus t_k$ has a stronger distinguishing ability than t_j or t_k.

Under the selected the corpus C and the core keywords, The term-matrix **C** is uniquely determined by the corpus C and the core keywords except for the order of the rows and columns. Therefore, $\mathrm{I}(t_j)$ is determined by C.

4 Thematic Membership Degree

Different keyword has the different semantic information measurement. Thus, in the same document, the contribution of different keywords to the thematic membership degree of the document is different.

Definition 6 Information Measure Distribution. Let C be some thematic corpus, $\mathrm{T_C}$ the set of all the core keywords. $F(\mathrm{C})=\{(t_j,\mathrm{I}(t_j))|\ t_j\in\mathrm{T_C},\ j=1,\cdots,n\}$, Call $F(\mathrm{C})$ the information measure distribution of the corpus C.

$F(\mathrm{C})$ is uniquely determined by corpus.

Definition 7 Thematic Similarity Degree. Let $\mathrm{T_{C1}}$ and $\mathrm{T_{C2}}$ be the two sets of core keywords of the themes C_1 and C_2, $\mathrm{I_{C1}}(t_j)$ and $\mathrm{I_{C2}}(t_j)$ be the information measures of the t_j in the themes C_1 and C_2 respectively. $S(\mathrm{C}_1,\mathrm{C}_2)= \dfrac{2\sum_{t_j\in \mathrm{T_{C1}}\cap\mathrm{T_{C2}}}\min(\mathrm{I_{C1}}(t_j),\mathrm{I_{C2}}(t_j))}{\sum_{t_j\in\mathrm{T_{C1}}}\mathrm{I_{C1}}(t_j)+\sum_{t_j\in\mathrm{T_{C2}}}\mathrm{I_{C2}}(t_j)}$,

Call $S(\mathrm{C}_1,\mathrm{C}_2)$ the thematic similarity degree of C_1 and C_2.

Obviously, $0\le S(\mathrm{C}_1,\mathrm{C}_2)\le 1$ for any two themes C_1 and C_2.

C_1 and C_2 are said to be the same or approximately similar if there exists a small threshold value $\varepsilon\ge 0$ such that $S(\mathrm{C}_1,\mathrm{C}_2)\ge 1-\varepsilon$.

If C_1 and C_2 are the same or approximately similar, then $F(\mathrm{C}_1)\approx F(\mathrm{C}_2)$.

Let $\mathrm{T}=\{t_1^{(i)},\cdots,t_{ki}^{(i)}\}$ be the set of all the keywords of the document d_i($i=1,\cdots,m$). Might as well supposition $t_1^{(i)},\cdots,t_{pi}^{(i)}$ ($p_i<k_i$) are core keywords.

$$S_1(\mathrm{C})=\max_{1\le i\le m}\sum_{j=1}^{p_i}\mathrm{I}(t_j^{(i)})\,,\ S_2(\mathrm{C})=\ \max_{1\le i\le m}\sum_{j=1}^{p_i}(f_i(t_j^{(i)})\cdot\mathrm{I}(t_j^{(i)}))\,.$$

Definition 8 Document Information Measure. For any document X, let T_X be the set of all the keywords of X, $f_X(t)$ be the word frequency of t, among them $\mathrm{T}_{X\cap\mathrm{C}}$ is the set of all the core keywords. $S_1(\mathrm{C}|X) = \sum_{t\in T_{X\cap C}}\mathrm{I}(t)$, $S_2(\mathrm{C}|X)= \sum_{t\in T_{X\cap C}}(f_X(t)\cdot\mathrm{I}(t))$, call $S_1(\mathrm{C}|X)$ and $S_2(\mathrm{C}|X)$ the first and second information measure of X based on theme C respectively.

Definition 9 Document Membership Degree. $X(t_1,\cdots,t_l)$, T_X, $f_X(t_1)$,$\cdots$, $f_X(t_l)$, C, $\mathrm{T}_{X\cap\mathrm{C}}$, $S_1(\mathrm{C}|X)$, $S_1(\mathrm{C})$, $S_2(\mathrm{C}|X)$, $S_2(\mathrm{C})$ are the same as above, $P(\mathrm{C}|X)= |\mathrm{T}_{X\cap\mathrm{C}}|/|T_X|$, $\rho_1(\mathrm{C}|\mathrm{X})= P(\mathrm{C}|X)\ S_1(\mathrm{C}|X)/\ S_1(\mathrm{C})$, $\rho_2(\mathrm{C}|\mathrm{X})= P(\mathrm{C}|X)\ S_2(\mathrm{C}|X)/\ S_2(\mathrm{C})$, call $\rho_1(\mathrm{C}|X)$, $\rho_2(\mathrm{C}|X)$ are the first and second membership degree of X belonging to theme C respectively.

The definition of the document membership degree shows that the more core keywords, the bigger the ρ_1(C|X) and ρ_2(C|X). The bigger the information measure, the bigger the ρ_1(C|X) and ρ_2(C|X).

Document classification is very important content in knowledge discovery.

Suppose that the information measure distributions of theme $C_1,\cdots,C_m$ are $F(C_1),\cdots,F(C_m)$ respectively. For any document X, it is evaluated as belonging to the theme C_i, which makes $\rho_2(C_i|X)$ maximum.

5 Evaluation

The empirical data come from the open platform for CNLP (http://www.nlp.org.cn/docs/doclist.php?cat_id=16&type=15). There are 19637 documents, divided into 20 categories in the data set. Training / test documents ratios are basically 1:1.

Due to the limited space, The following part only introduce the results of the experiment of the theme “Computer”.

There are 1024 documents in the computer corpus. Select only the top 600 documents as the training corpus. After word segmenting, a total of 7199 keywords are obtained.

The core keywords set of the computer corpus is defined as: $T_C = \{t_j \mid \sum_{i=1}^{m}[f_i(t_j)\cdot(\log m - \log f(t_j)) > 250\}$ = { "B超", "IP组", "Petri网" , "一切" , "一致性" , "一起" , "上下文" ,……}, totally come to 1055.

F(C) = {("B超", 0.5000), ("IP组", 0.1170), ("Petri网", 0.6957), ("一切", 0.4048), ("一致性", 0.6401), ("一起", 0.1992), ("上下文", 0.3179), ……}.

The results of the performance of the ρ_1(C|X) and ρ_2(C|X) are shown in table 1. The threshold values of ρ_1(C|X) and ρ_2(C|X) are set at 0.2236 and 0.0680 when 90% of all training documents are determined to belong to the training theme “Computer”.

Table 1. Test results of document Membership Degree

Tested theme	Number	ρ_1 avg.	ρ_1>0.2236	ρ_2 avg.	ρ_2>0.0680
C3-Art	50	0.1183	0.00%	0.0410	12.00%
C4-Literature	26	0.0458	0.00%	0.0066	0.00%
C5-Education	50	0.0456	0.00%	0.0079	0.00%
C6-Philosophy	45	0.0680	0.00%	0.0146	0.00%
C7-History	50	0.1194	2.00%	0.0446	20.00%
C11-Space	49	0.2607	59.18%	0.1046	59.18%
C15-Energy	33	0.0361	0.00%	0.0047	0.00%
C16-Electronics	28	0.0692	0.00%	0.0116	0.00%
C17-Communication	27	0.0594	0.00%	0.0100	0.00%
C19-Computer	50	0.3935	96.00%	0.2001	94.00%
C23-Mine	34	0.0296	0.00%	0.0042	0.00%
C29-Transport	50	0.0276	0.00%	0.0035	0.00%
C31-Enviornment	50	0.1757	22.00%	0.0589	40.00%
C32-Agriculture	50	0.1744	14.00%	0.0553	24.00%
C34-Economy	50	0.1796	22.00%	0.0654	42.00%
C35-Law	50	0.0427	0.00%	0.0103	0.00%
C36-Medical	50	0.0212	0.00%	0.0026	0.00%
C37-Military	50	0.0209	0.00%	0.0034	0.00%
C38-Politics	50	0.1539	14.00%	0.0607	30.00%
C39-Sports	50	0.1683	18.00%	0.0523	20.00%

The testing results show that 12% documents of the theme C3 are determined to belong to the training theme C (Computer) with the set threshold values. This is because the theme C3 is similar to the training theme C in some extent. The theme C7, C11 etc are also similar to the training theme C in some different extent.

6 Conclusions and Future Work

This paper defines the keyword information measure based on the corpus. Experiments show that the corpus-based information measure of the key words and the membership degree of the documents are effective and have stronger ability to distinguish different themes of documents.

In future work, we plan to train more different theme corpora to get their corresponding distributions of information measure, and establish a backup information measure distribution database for real applications.

References

[1] Lu, R.-Q.: Forefront of scientific knowledge and research. China Awards for Science and Technology 8(4) (2000)

[2] Li, L.-F., Tan, J.-r., Liu, B.: Quantitative information measurement and application for machine component classification codes. Journal of Zhejiang University Science 6A(suppl. I), 35–40 (2005)

[3] Geogre, J.K.: An update on generalized information theory. In: ISIPTA, pp. 321–334 (2003)

[4] Weaver, W.: Recent Contributions to the Mathematical Theory of Communication. Mathematical Theory of Communication. University of Illinois Press, Urbana (1949)

[5] Bar-Hillel, Y., Carnap, R.: An outline of a theory of semantic information. Tech. Rep. No. 247. Research Lab. of Electronics. MIT, Cambridge (1952)

[6] Zadeh, L.A.: Fuzzy Sets. Information Control 8, 338–353 (1965)

[7] Pawlak, Z.I.: Rough sets. International Journal of Computer and Information Sciences (11), 341–356 (1982)

[8] Deerwester, S., Dumais, S.T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41, 391–407 (1990)

[9] Chen, N., Chen, A., Zhou, L.X.: A Documental Clustering Algorithm Based on Fuzzy Concept Graph and Its Application in WebMEIHPGCH. Journal of Software 13(8) (2002)

[10] Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998)

[11] Hofmann, T.: Probabilistic Latent Semantic Indexing. In: 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval, Berkeley, California, pp. 50–57 (1999)

[12] He, M., Feng, B., Fu, X.: Web Document Classification Based on Rough Set Latent Semantic Indexing. Computer Engineering 30(13) (2004)

[13] He, W.: LSI Latent Semantic Indexing Model. Mathematics in Practice and Theory 33(9) (September 2003)

[14] Zhou, S.-g., Guan, J.-h., Hu, Y.-f.: Latent Semantic Indexing (LSI) and its Applications in Chinese Text Processing. Mini-Micro System 22(2) (February 2001)

[15] Manning, C.D., Schäutze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)

A New Similarity Measure Based Robust Possibilistic C-Means Clustering Algorithm

Kexin Jia, Miao He, and Ting Cheng

School of Electronic Engineering, University of Electronic Science and Technology,
Chengdu, 611731, China
Jiakexin@sina.com

Abstract. In this paper, we focus on the development of a new similarity measure based robust possibilistic c-means clustering (RPCM) algorithm which is not sensitive to the selection of initial parameters, robust to noise and outliers, and able to automatically determine the number of clusters. The proposed algorithm is based on an objective function of PCM which can be regarded as special case of similarity based robust clustering algorithms. Several simulations, including artificial and benchmark data sets, are conducted to demonstrate the effectiveness of the proposed algorithm.

Keywords: Fuzzy c-means clustering, Possibilistic c-means clustering, Automatic merging, Robustness, Noise and outliers.

1 Introduction

Cluster analysis is a method for finding groups within data with most similar objects in the same cluster and most dissimilar objects in different clusters. Since Zadeh[1] proposed fuzzy sets that produced the idea of partial memberships described by a membership function, fuzzy clustering has been widely studied and applied in a variety of key areas [2-5]. In fuzzy clustering, the fuzzy c-means (FCM) clustering algorithm, firstly proposed by Dunn [6] and then extended by Bezdek [2], is the best known and used method. Although FCM is very useful clustering algorithm, its membership does not always correspond well to the membership degree of data point. It may be inaccurate in a noisy environment.

To improve this weakness of FCM, and to produce memberships that have good explanation of the membership degree for data points, Krishnapuram and Keller [7] relaxed the constraint on the sum of memberships to create a possibilistic approach to clustering where they called it a possibilistic c-means (PCM). Krishnapuram and Keller showed that algorithms with possibilistic c-memberships are more robust to noise and outliers than FCM. However, Barni et al. [8] presented that PCM may sometimes produce coincident clusters. Afterwards, Krishnapuram and Keller [9] provided more insights and recommendations where they made a conclusion that PCM can be seen as a mode-seeking algorithm. to improve PCM, several researchers such as Zhang and Leung[10], Pal and Keller [11] proposed a combined way of PCM with FCM. However, these PCM algorithms still face parameter selection, initialization, and

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 335–342, 2011.

cluster number determination. Vincent et al. [12] combined PCM and mountain method together, proposed a similarity based PCM algorithm which selected the optimal parameters by experiments. By embedding a robust merging algorithm into a modified PCM algorithm, Yang and Lai [13] proposed a robust automatic merging possibilistic clustering (AM-PCM) algorithm which considered all data points as initial cluster centers and discussed the robustness of the AMPCM algorithm to noise and outliers by simulation. The quantitative analysis of its robustness is not given.

In this paper, we firstly point out that the objective function used in [13] can be regarded as a special case of the total similarity measure defined in [14], and then propose a new special case of the total similarity measure which corresponds to a new objective function of PCM algorithms. Combining the objective function and the automatic merging approach discussed in [13] together, a new similarity measure based RPCM algorithms is proposed.

2 Objective Function of AM-PCM Algorithm

Let $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ be a set of N data points in s-dimensional space where the jth data point $\boldsymbol{x}_j^T=\left[x_{jd}\right]_{1\times s}$. Clustering analysis is a technique used to partition the data set $\boldsymbol{X}$ into C subsets that can well represent the data structure of $\boldsymbol{X}$. The partition of C clusters can be described by a possibilistic membership matrix $\boldsymbol{U}=\{\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_C\}=\{u_{ij}\}_{N\times C}$, where u_{ij} represents the possibilistic membership of $\boldsymbol{x}_i$ belonging to the jth cluster.

The objective function used in [13] is given as follows:

$$J_1(\boldsymbol{U},\boldsymbol{V})=\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}d_{ij}^2-\eta\sum_{i=1}^{C}\sum_{j=1}^{N}u_{ij}\left(1-\frac{\gamma}{\gamma+1}u_{ij}^{1/\gamma}\right) \tag{1}$$

where $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_C\}$ is the set of C cluster centers; d_{ij} is a dissimilarity measure where $d_{ij}=d(\boldsymbol{x}_j,\boldsymbol{v}_i)=\|\boldsymbol{x}_j-\boldsymbol{v}_i\|$ is the Euclidean distance between $\boldsymbol{x}_j$ and $\boldsymbol{v}_i$; $\gamma>0$, and η is the maximum squared Euclidean distance of the data points in $\boldsymbol{X}$.

By differentiating (1) with respect to u_{ij} and $\boldsymbol{v}_i$, respectively, and set them to 0, we obtain the update equations for u_{ij} and $\boldsymbol{v}_i$ as follows

$$u_{ij}=\left(\frac{\eta-d_{ij}^2}{\eta}\right)^{\gamma},\ \boldsymbol{v}_i=\frac{\sum_{j=1}^{N}u_{ij}\boldsymbol{x}_j}{\sum_{j=1}^{N}u_{ij}} \tag{2}$$

From the above construction, only the parameter γ is remained and necessary to be estimated. In [13], the equation of estimating γ is as follows:

$$\gamma = \frac{\log\left(1-\frac{1}{p}\right)}{\log\left(\frac{\eta - D}{\eta}\right)} \tag{3}$$

where p is chosen between 3 and 4; D can be obtained by using these $N(N-1)/2$ squared Euclidean distances of all data points. The detailed derivation of D can be found in [13].

3 Proposed RPCM Algorithms

In this section, we firstly discuss the relationship between the objective function used in [13] and the total similarity measure defined in [14], and then proposes a new RPCM algorithm.

Substituting (2) and into (1), the objective function can be rewritten as

$$J_1(\boldsymbol{U},\boldsymbol{V}) = -\frac{\eta}{\gamma+1}\sum_{i=1}^{C}\sum_{j=1}^{N}\left(\frac{\eta - d_{ij}^2}{\eta}\right)^{\gamma+1} \tag{4}$$

Hence, minimizing the above objective function can be regard as maximizing the following similarity measure function

$$J_{s1}(\boldsymbol{V}) = \frac{\eta}{\gamma+1}\sum_{i=1}^{C}\sum_{j=1}^{N}\left(\frac{\eta - d_{ij}^2}{\eta}\right)^{\gamma+1} \tag{5}$$

For fixed γ, the above equation can be regarded as a special cases of the following robust similarity measure discussed in [14].

$$J_s(\boldsymbol{V}) = K\cdot\sum_{i=1}^{C}\sum_{j=1}^{N}\left[S\left(d_{ij}^2\right)\right]^{\lambda} \tag{6}$$

where $\lambda > 0$ and $S(\cdot)$ is a monotone increasing similarity measure function with the decrease of squared Euclidean distance d_{ij}^2. The constant K is introduced to simplify the update equations of u_{ij} and v_i. When we set $K = \eta/(\gamma+1)$, $\lambda = \gamma+1$ and $S\left(d_{ij}^2\right) = \left(\eta - d_{ij}^2\right)/\eta$, the equation (5) can be found.

Base on equation (6), we will define a new similarity measure function which corresponds to Gaussian similarity measure based objective function. Substituting $K = \eta/\gamma$, $\lambda = \gamma$, and Gaussian similarity measure $S\left(d_{ij}^2\right) = \exp\left(-d_{ij}^2/\eta\right)$ into (6), the total similarity measure function can be given by

$$J_{s2}(\boldsymbol{V}) = \frac{\eta}{\gamma}\sum_{i=1}^{C}\sum_{j=1}^{N}\left\{\exp\left(-\frac{d_{ij}^2}{\eta}\right)\right\}^{\gamma} \tag{7}$$

which corresponds to the objective function

$$J_2(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij} d_{ij}^2 + \frac{\eta}{\gamma}\sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}\left(\log u_{ij} - 1\right) \tag{8}$$

where $\gamma > 0$, and η is the estimated sample variance, that is, η could be defined by

$$\eta = \frac{1}{N}\sum_{j=1}^{N}\left\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\right\|^2, \quad \overline{\boldsymbol{x}} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}_j \tag{9}$$

By differentiating (8) with respect to u_{ij} and $\boldsymbol{v}_i$, respectively, and set them to 0, we obtain the update equations for u_{ij} and $\boldsymbol{v}_i$ as follows

$$u_{ij} = \left\{\exp\left(-\frac{d_{ij}^2}{\eta}\right)\right\}^{\gamma}, \quad \boldsymbol{v}_i = \frac{\sum_{j=1}^{N} u_{ij}\boldsymbol{x}_j}{\sum_{j=1}^{N} u_{ij}} \tag{10}$$

The equivalence of (7) and (8) can be easily proved by substituting (10) into (8).

Based on (10), the parameter γ can be estimated by using the same derivation discussed in [13]. The equation of estimating γ is as follows:

$$\gamma = \frac{-\eta \log\left(1 - \frac{1}{p}\right)}{D} \tag{11}$$

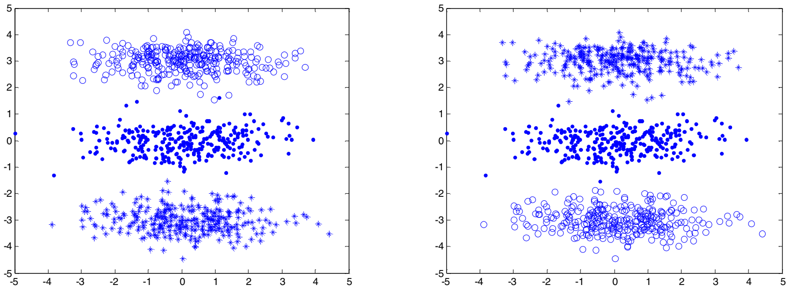
where the parameters p and D is defined as these in [13].

Combining the automatic merging approach discussed in [13] and the equations (10) and (11), a Gaussian similarity measure based robust possibilistic clustering (GRPCM) algorithm can be created.

4 Experimental Examples

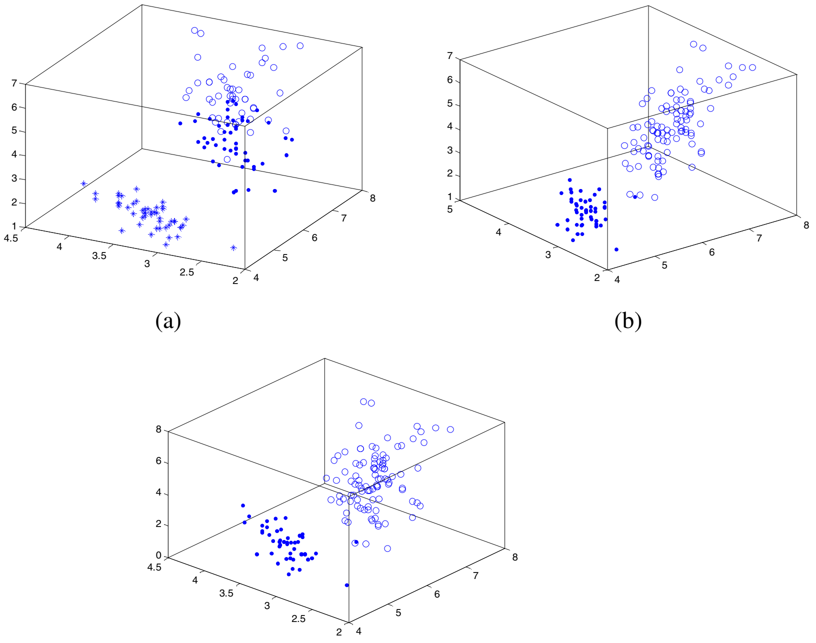
In this section, we use some experimental examples to demonstrate effectiveness and superiority of the proposed GRPCM algorithm. In all examples, we give $p = 3$, $\rho = 0.9$, and $\varepsilon = 10^{-3}$.

Example 1: The data set as shown in Fig.1 (a) is generated from a three-component Gaussian mixture distribution where each component has 300 data points. We use this data set to illustrate behaviors of GRPCM under different running steps. We start implementing GRPCM with all data points as initial cluster centers. After the first iteration, some clusters are merged such that the cluster number is reduced from 900 to 215. After the 61st iteration, the GRPCM converges with 3 identified clusters. The final 3 identified clusters are shown in Fig. 1(b). In the procedure, we can find that the cluster number decreases quickly when it merges those data points with coincident cluster centers.



(a) (b)

Fig. 1. (a) the data set used in Example 1; (b) the final identified three clusters from GRPCM

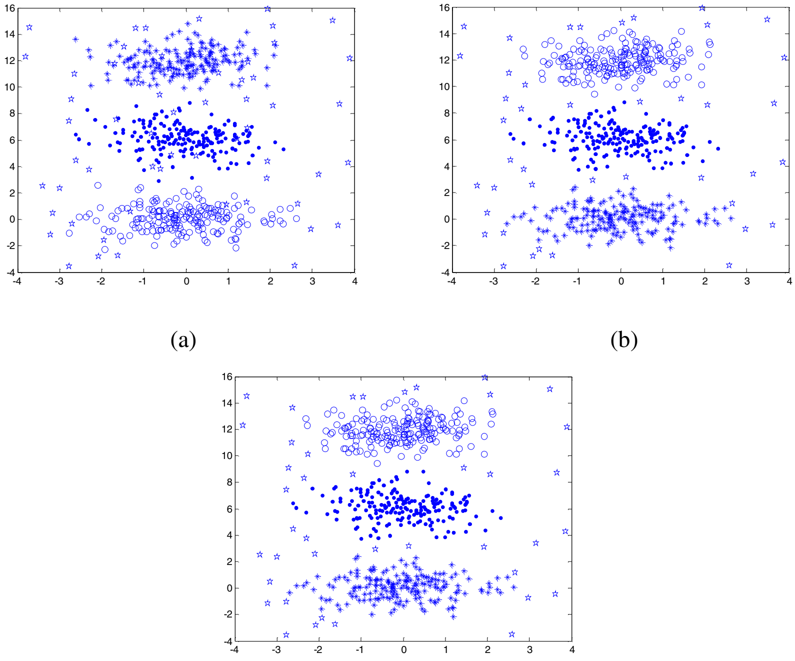


(a) (b)

(c)

Fig. 2. (a) Iris data set used in Example 2; (b) AMPCM; (c) GRPCM

Example 2: This example uses Iris data as a test set that is a benchmark data set frequently used in the pattern recognition literature. The Iris data set consists of four input measurement: sepal length, sepal width, petal length, and petal width. Three species of Iris are involved and each kind of Iris contains 50 instances. Fig.2 (a) shows the Iris data projected into the axes of the first, second, and third measurements. The optimal cluster number of clusters is 2 or 3, because there exists two overlapped clusters. Suppose that we do not know what the cluster number is. We try to use cluster validity indexes to find optimal cluster number estimates by using FS[16], XB[17], AMB[18], Kwon[19], Tang[20], PCAES[21], FPBM[22], NZ[23], and CY[24]. The values of FS, XB, AMB, Kwon, Tang, PCAES, FPBM, NZ, and CY are shown in Table 1. The indexes FS, XB, Kwon, Tang and CY give optimal cluster number estimates $C = 2$ or 3, but the index FPBM gives an optimal cluster number estimate $C = 9$, the index PCAES gives an optimal cluster number estimate $C = 5$ and the indexes NZ and AMB give an optimal cluster number estimate $C = 10$. As shown in Fig. 2 (b) and (c), the AMPCM and GRPCM algorithms give optimal cluster number estimate $C = 2$, and can not process overlapped clusters.



(a)

(b)

(c)

Fig. 3. (a) The data set used in Example 3; (b) AMPCM; (c) GRPCM

Example 3: The data set as shown in Fig.3 (a) includes four clusters. The first three clusters are generated from a three-component Gaussian mixture distribution where each component has 180 data points and the fourth cluster is an outlier set generated from a uniform rectangle-shape data set with sample sizes 60. The identified noise and outliers are marked with pentagrams. The cluster results of the above mentioned two algorithms are shown in Fig. 3(b) and (c). It can be seen that the proposed algorithms are robust to noise and outliers.

Table 1. Values of nine cluster validity indexes

Q	FS	XB	AMB	Kwon	Tang	PCAES	FPBM	NZ	CY
2	-370.6	0.0542	0.1018	8.387	8.857	2.600	0.2267	-8.077	0.0119
3	-443.7	0.1371	0.0580	21.98	22.29	2.707	0.3503	-0.747	0.0642
4	-432.1	0.1957	0.0446	32.04	31.04	4.164	0.3823	-4.449	0.0513
5	-374.5	0.3997	0.0265	68.31	59.22	1.961	0.3619	0.910	0.0017
6	-381.6	0.3016	0.0205	54.23	51.17	7.443	0.3670	3.147	0.0016
7	-342.6	0.5435	0.0157	100.64	82.67	6.354	0.4113	3.385	0.0012
8	-335.4	0.4406	0.0143	84.16	73.65	6.092	0.3990	3.348	0.0011
9	-321.0	0.3836	0.0132	75.24	69.18	7.164	0.4145	-2.971	0.0016
10	-299.3	0.6620	0.0107	134.62	105.6	5.583	0.4116	3.535	0.0016

5 Conclusion

This paper proposes a new objective function of PCM whose relationship with the one discussed in [13] is described in detail. All these objective functions can be regarded as special cases of the robust similarity based clustering method studied in [14]. Combining the objective function and automatic merging approach together, this paper also proposed a new robust PCM algorithms-GRPCM. Simulation results show that the proposed algorithm and the AMPCM have very similar performance. In addition, the proposed algorithm can use the same processing method as AMPCM algorithm for very large data sets.

References

1. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
2. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithm. Plenum Press, New York (1981)
3. Yang, M.S.: A survey of fuzzy clustering. Mathematical and Computer Modeling 18, 1–16 (1993)
4. Baraldi, A., Blonda, P.: A survey of fuzzy clustering algorithms for pattern recognition-part I and II. IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics 29, 778–801 (1999)
5. Hoppner, F., Klawonn, F., Kruse, R.: Fuzzy cluster analysis: methods for classification data analysis and image recognition. Wiley, New York (1999)
6. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters. Journal of Cybernetics 3, 32–57 (1974)
7. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1, 93–110 (1993)
8. Barni, M., Cappellini, V., Mecocci, A.: Comments on 'A possibilistic approach to clustering'. IEEE Transactions on Fuzzy Systems 4, 393–396 (1996)
9. Krishnapuram, R., Keller, J.M.: The possibilistic C means algorithm: Insights and recommendations. IEEE Transactions on Fuzzy Systems 4, 385–393 (1996)
10. Zhang, J.S., Leung, Y.W.: Improved possibilistic C-means clustering algorithms. IEEE Transactions on Fuzzy Systems 12, 206–217 (2004)

11. Pal, N.R., Pal, K., Keller, J.M., Bezdek, J.C.: A possibilistic fuzzy c-means clustering algorithm. IEEE Transactions on Fuzzy Systems 13, 517–530 (2005)
12. Tseng, V.S., Kao, C.P.: A novel similarity-based fuzzy clustering algorithm by integrating PCM and mountain method. IEEE Transactions on Fuzzy Systems 15(6), 1188–1196 (2007)
13. Yang, M.S., Lai, C.Y.: A robust automatic merging possibilistic clustering method. IEEE Transactions on Fuzzy Systems 19(1), 26–41 (2011)
14. Yang, M.S., Wu, K.L.: A similarity-based robust clustering method. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(4), 434–448 (2004)
15. Huber, P.J.: Robust statistics. Wiley, Chichester (1981)
16. Fukuyama, Y., Sugeno, M.: A new method of choosing the number of clusters for fuzzy c-means method. In: Proceeding of Fifth Fuzzy Systems Symposium, pp. 247–250 (1989)
17. Maulik, U., Bandyopadhyay, S.: Performance evaluation of some clustering algorithm and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(12), 1650–1654 (2002)
18. Bensaid, A.M.: Validity-guided (re)clustering with applications to image segmentation. IEEE Transactions on Fuzzy Systems 4(2), 112–123 (1996)
19. Kwon, S.H.: Cluster validity index for fuzzy clustering. Electronic Letters 34(22), 2176–2177 (1998)
20. Tang, Y.G., Sun, F.C., Sun, Z.Q.: Improved validation index for fuzzy clustering. In: American Control Conference, Portland, USA, June 8-10 (2005)
21. Wu, K.L., Yang, M.S.: A validity index for fuzzy clustering. Pattern Recognition Letters 26, 1275–1291 (2005)
22. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern Recognition 37, 481–501 (2004)
23. Zahid, N., Limouri, M., Essaid, A.: A new cluster-validity for fuzzy cluster. Pattern Recognition 32, 1089–1097 (1999)
24. Cho, S.B., Yoo, S.H.: Fuzzy Bayesian validation for cluster analysis of yeast cell-cycle data. Pattern Recognition (2005)
25. Lubischew, A.A.: On the use of discriminant functions in taxonomy. Biometrics 18, 455–477 (1962)

DOM Semantic Expansion-Based Extraction of Topical Information from Web Pages

Junjie Chen[*], Junyao Jia, and Liguo Duan

Computer Science & Technology College
Taiyuan University of Technology
Taiyuan, China
chenjj@tyut.edu.cn,
283649126@sina.com

Abstract. Web pages usually contain much irrelevant information that customers don't need. Thus, in order to extract relevant information from the complicated information heap, effective methods to extract information are required. Aiming at the semi-structured characteristic of HTML, theme-relevant information in web pages could be extracted by semantic pruning, in the adoption of DOM-presentation, combined with the feature of web structure and the fuzzy classification of keywords.

Keywords: DOM Tree, Information extraction, Partition, Semantic expansion, Correlativity, Fuzzy classification.

1 Introduction

The rapid development of Internet greatly enriched the number of information on the Internet. All kinds of information glut in our information word, exactly to the worlds in every page contain much "noise information" such as navigation and the advertising messages, pictures, etc. At the same time, it bring us a question, how to extract the WEB theme information from the local downloaded WEB pages that extracted from the open-domain question answering system. This information extraction technology has become one of the popular research fields.

2 Related Research

A great deal of research work about Web information extraction has been done at home and abroad.

Literature [2] combined the semantic tree in STU described with DOM tree, and the STU - DOM model will convert HTML to STU –DOM Tree, meanwhile it introduced the partial relevancy and context relevancy to filter and prune the DOM tree, and get the information related to the theme. The method only considered the partial relevancy and the context relevancy. It has certain limitations.

[*] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 343–350, 2011.

Literature [3] completed the conversion from HTML page to the DOM tree, considering the influence of label type and the tag of DOM tree to the theme, expanded the node and calculated its semantic effective factor and compared to extract Web pages in order to prune the tree and get the information content. But the method emphasize labels and tags attributes and it is not suitable to the WEB page with no significant tag characteristics.

Literature [4] filtered the irrelevant content by setting filter structure. But in the progress of deleting the uncorrelated content with the theme, without considering the statements context relevancy, the useful information may be deleted.

Literature [5] put forward the extraction method based on parallel fuzzy classification of web pages, using corresponding words to replace the web page words, accelerate the extraction rate and raise the extraction accuracy. In fact, the HTML structured was not extracted in the method, thus the extraction accuracy is not high.

Refer to the above theme information acquisition methods, combined with the statistic and observation of HTML pages, this paper proposed a new extraction algorithm of web page subject information. This algorithm is the characteristic of structure, relying on DOM label type and comprehensive consideration of the effective factors of node properties, and the fuzzy matching method. It does not delete the useful node. The experimental results prove that the theme information method has high accuracy and integrity.

3 Algorithm Described

3.1 Basic Definition

Definition 1: DOM (Document Object Model), is the standard interface formulated W3C specification. Web pages are usually transformed into DOM tree via HtmlParser, for example figure1. Every node in the tree is a object. DOM model shows the document structure, and defines the behavior of the node object. According to the method and attribute of the node object, DOM tree node can be visited, modified, added and deleted.

Definition 2: Zadeh proposed the Fuzzy set theory in 1965. U stand for a data base. $U=\{u_1,u_2, \quad ,u_n\}$. A is a fuzzy set based on U, A can be expressed as:

$$A=\{(u_1,f_A(u_1)),(u_2,f_A(u_2)), \quad ,(u_n,f_A(u_n))\} \tag{1}$$

f_A, $f_A:U\rightarrow[0,1]$ is a subordinate function on fuzzy sets; $f_A(u_i)$ is the membership in A.

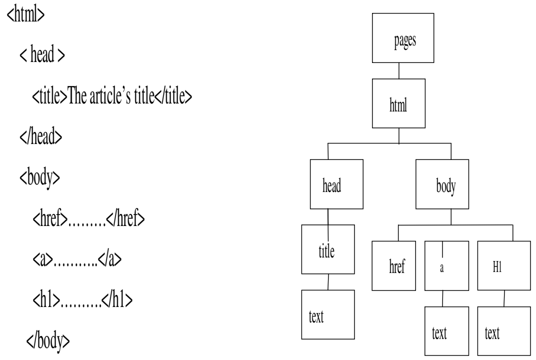
Definition 3: The effection node, show the relative influence on the content with Inf (node), Inf (node)∈ [0, 1]. The higher the value, the higher level of that affect.

The factors determined the effective node contain the node category and tag. Initial value go down follow title, content, vision , block , link and other. $Init(node_i)$

stand for the effective factors. And its value can be passed. Therefore, the effective value of leaf nodes under can be made in calculation:

$$Inf\ (leaf\ ) = \sum_{i=1}^{k} Inf\ (\ Ancestor\ _{i}) \tag{2}$$

$Inf\,(Ancestor_i)$ is ancestors node, k is the number of Ancestor nodes.



```
<html>
  < head >
    <title>The article's title</title>
  </head>
  <body>
    <href>.........</href>
    <a>...........</a>
    <h1>..........</h1>
  </body>
```

Fig. 1. HTML source code and DOM tree

3.2 Algorithm Review

The extraction Web page subject information system is divided into six parts: the HTML parsing, filtering, fuzzy classification, semantic analysis, semantic expansion and pruned. As shown in figure 2 is frame-chart information extraction.

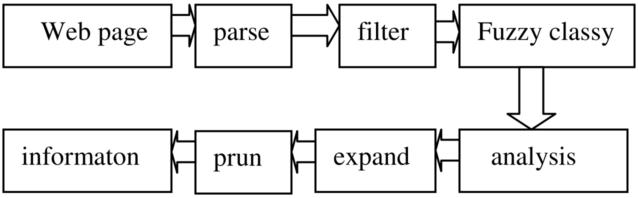


Fig. 2. Information extraction frame-chart

HTML analyzing is transforming Web pages into the DOM tree used by HTML Parser. The second step is to filter the nodes has nothing to do with in the DOM tree, such as pictures < I mg >, links < href >, table<form > and the HTML tags. Delete all these nodes, and if no found, return.

Fuzzy classification is to use the corresponded words stored in words base to replace the web words, making the expressions of web information obtained normalized. Filtered DOM tree no longer have pictures and link information, and contain a large amounts of information in words. Fuzzy classification technology can bring q&a system problem analysis extracted keywords make trade-offs. Because a lot of web information show in different forms but same meaning words, such as "计算机" and "电脑". These two words have same meaning, can set them for "计算机" in unified description.

As to the words not covered, the choice is to choose the most co-occurrence rate word as its concept. The words with many semantics, such as "软件", can be labeled as "computer software" and "social environment", etc. Choosing the concept tagging appeared many times. This system adopt fuzzy set theory give the web pages a information description, make the text extraction more easier and accurate.

Semantic analysis is add semantic attribute value to the DOM tree node, including two types of values, the partial relevancy and text relevancy. Partial correlation is determined by the unlink Chinese characters in block and the total amounts of links. The calculation formula can be expressed as:

$$local\,(block_i) = \frac{words\,(block_i)}{links\,(block_i)} \tag{3}$$

$$links\,(block_i) = \sum_{j=1}^{N} links\,(block_{ij}) \tag{4}$$

$$words\,(block_i) = \sum_{j=1}^{N} words\,(block_{ij}) \tag{5}$$

$block_{ij}$ is the j of the $block_i$ sub-tree. $links(block_i)$ is the link property values of $block_i$. $words(block_i)$ is the word property values of $block_i$. If $local(block_i) \geq L$ (L is partial relevancy threshold), then the node local related. The context relevancy is the correlation judgment about father nodes and child nodes of local relevance, only when all three local related, can judge the node is associated with theme information.

DOM tree semantic expand take advantage of the influence of theme information from HTML tags categories and labels, and add the influence to the relevancy computation. The node use node effective factor depicts the important degree. Web pages are information carrier, and show the grounds with markers of discrete text composed of string. The markers control the sequence of information, decided how to display definition of the text, pictures and etc. In order to increase the DOM tree node and the related page semantic theme information, HTML tags categories (Category), the link text (WordNum) and the effect of Influence factors (Influence) attribute value are added to expand its semantic. According to the correlation, HTML tags can be divided into five types: (1) describing the title and page summary information , such as < title > , < meta >, etc. (2) planning web page layout label, such as< table>, < tr >, < td >, < p >, < div >, etc. (3) describing the display characteristics of key content , such as the labels, < b >, < I >, < strong > , <h>, tec. (4) hyperlinks related tags. (5) other tags(in text extraction will ignore such labels). Above 5 kinds of page subject important tags in turn decrease.

The main information of Web page contained in the most of the DOM tree leaf nodes, the other DOM nodes are mainly used for partition and display the appearance characteristics. In the existing page information extraction methods, for these people only consider content block action, and ignored node can reflect the important degree by characteristics of text label. For example, properties of < b > < font > tags. In addition, different classes node has different effect on its sons. For example, the block with a title for the ancestors nodes, the content of the block importance should be higher.

The pruning device judge the DOM tree node's local relevancy, the effective factor, the context relatedness accumulative get through practice in advance, and numerical comparison, the threshold value obtained from tree which deleted the small correlation of nodes information, finally small web page subject information was output.

The pruning device judge the DOM tree node's local relevancy, the effective factor, the context relatedness accumulative get through practice in advance, and numerical comparison, the threshold value obtained from tree which deleted the small correlation of nodes information, finally small web page subject information was output.

3.3 Fuzzy Classification Algorithm

Set $X = \{X_1, X_2, X_3 \quad , X_n\}$ to stay classification of the collections of objects, each object X_i stand for a set of characteristic data. Firstly, formula (1) is to explain X_i and X_j distance, and to establish fuzzy distance matrix.

$$S_{ij}^{f} = \sum_{k=1}^{m} \frac{|x_{ik} - x_{jk}|}{m} \tag{6}$$

For every X_i ∈ X, according to the given threshold to build similar $[X_i]_R$.

$$[X_i]_R = \{X_j | X_j \in X, S_{ij}^f \le A\} \tag{7}$$

Setting threshold for A:

$$A = 2 \times \sum_{i=1,j=1}^{i=n,j=n} \frac{S_{ij}^{f}}{n \times (n-1)} \tag{8}$$

For every X_j ∈ X, according to a given threshold build a similar set. Put S divided into category，if $S_{ij}^f \ge A$. Then merge the public elements with same class, and get the corresponded equivalent matrix. These equivalence classes are the results of clustered. Keywords contained in the Equivalence class can be interchanged without changing their semantic.

3.4 DOM Tree Expand Algorithm

DOM tree semantic expanded can divide into two steps: firstly, to traverse DOM tree, according to the effect factors of Init (node) to initial each node, identify the tag attributes and add the type; Then, for each leaf node in leaves set find its ancestors node bottom-up, accumulate the effective factor value of the ancestors , and calculate the value of the leaves nodes.

```
DOM Tree Expand(t)
  for  each node ∈t do
    node. Category={title| topvision| block| link| other}
    node. Attribution=[name, value]
    node. Influence=Init(node)
  end for
  value(leaves)
  for each leafnode∈leaves do
    sum=leafnode(i) The accumulative influence of
    ancestors nodes
    Leafnode. Influence=sum
  End for
end ‖DOM Tree Expand
```

4 Experimental Results and Analysis

In order to test the system's actual effect, the following experiment is designed. The question answering system in baidu-zhidao involves questions in education, culture, sports, entertainment and various fields (as shown in figure 3). Test system extracts the accuracy of answer source related to the theme. Experimental results are shown in table 1.

Fig. 3. Page examples

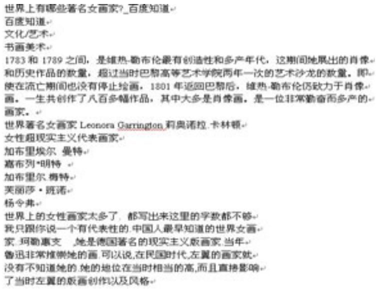
Artificial inspection analysis was carried out on 350 web pages in this experiment. The results indicate that a lot of web page noises are removed, theme information screening rate is high, so that web page subject information extraction effect is achieved. Figure 4 gives results of extraction examples in figure 3.

$$\text{Integrity} = \frac{\text{The number of completed theme informatio n pages}}{\text{The number of answer source pages}} \tag{9}$$

$$\text{compressibility} = \frac{\text{the size of result pages}}{\text{the size of answer source pages}} \tag{10}$$

Table 1. Experimental results

type	amount	integrity	compressibility
sports	100	94%	35%
education	50	96%	40%
culture	50	90%	46%
amusement	100	93%	30%
life	50	90%	44%



世界上有哪些著名女画家?_百度知道
百度知道
文化/艺术
书画美术
1783 和 1789 之间，是维热-勒布伦最有创造性和多产年代，这期间她展出的肖像
和历史作品的数量，超过当时巴黎高等艺术学院两年一次的艺术沙龙的数量。即
使在流亡期间也没有停止绘画，1801 年返回巴黎后，维热-勒布伦仍致力于肖像
画。一生共创作了八百多幅作品，其中大多是肖像画，是一位非常勤奋而多产的
画家。
世界著名女画家 Leonora Garrington 莉奥诺拉.卡林顿
女性超现实主义代表画家
加布里埃尔 曼特
嘉布列•明特
加布里尔.梅特
芙丽莎 • 妮诺
杨令弗
世界上的女性画家太多了，都写出来这里的字数都不够
我只跟你说一个有代表性的.中国人最早知道的世界女画
家.珂勒惠支 ,她是德国著名的现实主义版画家.当年
鲁迅非常推崇她的画.可以说,在民国时代,左翼的画家就
没有不知道她的.她的地位在当时相当的高,而且直接影响
了当时左翼的版画创作以及风格

Fig. 4. Instance of extraction results

5 Epilogue

This paper proposes a kind of automatic question answering system answers Web page subject information source method, by irrelevant information filtering through the DOM tree, fuzzy words classifying, local and context semantic analysis and calculation of nodes, semantic expansion of the effect factor, then cut the correlation based on node, in order to extract Web pages with relevant information. Experimental results indicate high integrity and compression ratio. However, this experiment was based on the answer library of the question answering system in baidu-zhida, in which the answer library Web information was limited. If this experiment broadens to all of the Web pages in the Internet, information acquisition coverage could be extended and the information could be extracted more completely, this method obtains a wider application.

References

1. Wang, T., Tang, S., Yang, D., et al.: COMIIX:Towards effective WEB information extraction, integration and query answering. In: Proc of SIGMOD 2002, p. 620. ACM Press, New York (2002)
2. Wang, Q., Tang, S., Yang, D., et al.: DOM-Based Automatic Extraction of Topical Information from Web Pages. Journal of Computer Research and Development 41(10), 1786–1791 (2004)

3. Gu, Y.-h., Tian, W.: Extraction of Information from Web Pages Based on Extended DOM Tree. Computer Science 36(11), 235–237 (2009)
4. Ou, J., Dong, S., Cai, B.: Topic information extraction from template web pages. Journal of Tsinghua University (Science and Technology) 45(1), 1743–1747 (2005)
5. Li, Y., Zhang, M.: A Fuzzy Extraction for Web Pages Based on Parallel Computing. Computer Engineering and Applications 21, 23–27 (2003)
6. Zheng, S.f., Liu, T., Qin, B., Li, S.: Overview of Question-Answering. Journal of Chinese Information Processing 16(06), 46–52 (2002)
7. Li, J.-j., Yan, H.-f.: Chinese Web Retrieval Test Collections: Construction, Analysis and Application. Journal of Chinese Information Processing (01), 30–36 (2008)
8. Gupta, S., Kaiser, G., Neistadt, D., Grimm, P.: DOM based Content Extraction of HTML Documents. In: 12th International World Wide Web Conference, vol. (5), pp. 207–214 (2003)
9. Freitag, D.: Machine learning for information extraction in information domains. Machine Learning 39(2/3), 169–202 (2000)

A Domain Specific Language for Interactive Enterprise Application Development

Jingang Zhou[1,2], Dazhe Zhao[1,2], and Jiren Liu[1,2]

[1] College of Information Science and Engineering, Northeastern University,
110004, Shenyang, China
[2] State Key Laboratory of Advanced Software Architecture (Neusoft Corporation),
110179, Shenyang, China
{zhou-jg,zhaodz,liujr}@neusoft.com

Abstract. Web-based enterprise applications (EAs) have become the mainstream for business systems; however, there are enormous challenges for EAs development to meet the software quality and delivery deadline. In this paper, we propose a domain specific language, called WL4EA, which combines components with generative reuse and targets for popular application frameworks (or platform) and supports high interactivity. With WL4EA, an EA can be declaratively specified as some sets of entities, views, business objects, and data access objects. Such language elements will be composed according to known EA architecture and patterns. Such a DSL and code generation can lower the development complexity and error proneness and improve efficiency.

Keywords: Domain Specific Language, Enterprise Application, Generative Programming, Web Application.

1 Introduction

Web-based application becomes the main stream of current enterprise application (EA) that deliveries business services through the web [1, 2, 3]. However, the development of such applications is still a complex system engineering and challenging, especially for those large ones [3, 4], e.g., budget overruns, timeline overdue, and unstable software, which even leads to projects cancelation. This is mainly due to two causes. One is unstable and implicitly requirements; the other is web application development is still an emerging and shifting paradigm [5, 6] that many established software engineering principles are not incorporated [6, 7, 8].

From the technological perspective, an EA typically involves many technologies, e.g., html/JavaScript/JSP/PHP, Java/C#, XML, SQL, etc. And usually the consistence among these technologies is not assured in the current development environment, which makes it hard to find a consistence bug [9]. Furthermore, the advent of Web 2.0 improves user experience, which appeals that an EA should have high interactivity and responsiveness like the desktop counterpart. Thus, Asynchronous JavaScript and XML (Ajax) becomes popular. Such a technical complexity makes it very difficult to develop an EA with quality.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 351–360, 2011.

Abstraction is the key to software construction [10], which encapsulates low level technology details with higher specifications or models. By using models, we can make software easy to understand and extend its life to a different (technological) platform with model transformation and refinement [11], which is the soul of model-driven development (MDD) [12]. As a lightweight and pragmatic MDD approach, domain-specific languages (DSLs) [13] got considerably adoption both in the academic and industry. Different from general purpose languages (GPLs, e.g., Java or C#), DSLs represent things in a specific domain with problem concepts, which lets developers perceive themselves as working directly with domain concepts rather than those of GPLs' if the DSL follows the domain abstractions and semantics as closely as possible. Such a domain orientation and abstraction can improve productivity greatly and lower cost of software development, which is evident from the example shown by SystemsForge that a web application (WA) can be constructed in a couple of days with DSLs and components [14] and other examples [15].

In this paper, we propose a DSL (actually a DSL family), called *WL4EA—**W**eb **L**anguage **f**or **E**nterprise **A**pplication*, for interactive EA development. WL4EA targets current main technology platform for EAs, i.e., the Java EE platform [16], and relies on some open source frameworks which are popular in EA community, as well as an Ajax library incorporated to support high interactivity for client computing. Our goal is to lower the technology threshold for EA developers and produce an EA with comparable quality to those written by skilled programmers.

The main contributions of this paper are:

- A DSL family for lightweight EA development.
- A DSL support high interactivity with the Ajax way.

The remainder of the paper is structured as follows. We discuss the design and technical detail of WL4EA in section 2. Related work is in section 3. We conclude this paper with future work in section 4.

2 WL4EA

A DSL is essentially composed of three components [15, 17], namely:

- *Abstract syntax*. The set of language and their relationships.
- *Concrete syntax*. The language notations used by end user to specify programs conforming to the abstract syntax.
- *Semantics*. The meaning of the language's constructs.

In the following subsections, we first show the technical and platform architecture for WL4EA to illustrate its semantics on a high level (the detail meaning of the elements are illustrated in the concrete syntax subsections of 2.3, 2.4, and 2.5), followed which we propose the metamodel of WL4EA for its abstract syntax, and finally discuss its concrete syntax in detail, as well as some implementation issues.

2.1 Technical and Platform Architecture

We target for interactive EAs, which are popular in recent years with the advent of Web 2.0. Therefore, we use Dojo, a famous Ajax-based open source framework with rich widgets and supported by many industry leaders—to implement the so called *rich* *internet applications* (RIAs) because Ajax is the main stream technology for realizing RIAs by our experience. We made some extensions and performance optimization on Dojo to make the UI widgets more intelligent and responsive. For example, we extended the original event mechanism of widgets listening for data validation for an efficient client computing, e.g., we rewrite the onBlur() for the controls so that we can validate the value of the control when the control is losing the focus. On the server side, we target for Java EE platform with lightweight open source frameworks—*Spring*, *Struts,* and *Hibernate (SSH).*

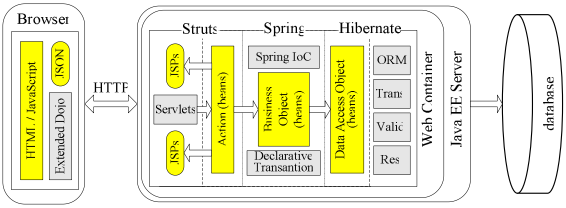


Fig. 1. Technical & platform architecture for WL4EA

Fig. 1 shows the target platform architecture, in which we make a clear layered style for an application, i.e., UI, business logic, and data access, thus, the goal of WL4EA is how to define the UI elements, actions, business objects, and data access objects backed by the Dojo and SSH. The elements with a yellow background in Fig. 1 are the target elements to be defined and generated, while the elements with a light gray background are the supporting techniques.

2.2 Metamodel

MDD provides a foundation to build DSLs by applying metamodeling. Metamodels provide a unified and expressive way to define the concepts of the domain. The use of metamodels to define abstract syntax helps reasoning about the domain and makes it possible interoperability with MDD tools [18].

A typical EA usually contains three layers, i.e., user interfaces (UI), business logic, and data persistence. And each contains some domain concepts. In WL4EA, we mainly model four categorized elements:

- *View*. The UI models elements which span client and web server covering page presentation with *theme* and *controls*, and *interactions* between client and server in the *Ajax* way through some controls' *events*.
- *Business Object* (BO). The model of business logic which concentrates on the static part of an EA, i.e., the skeleton of a service method.

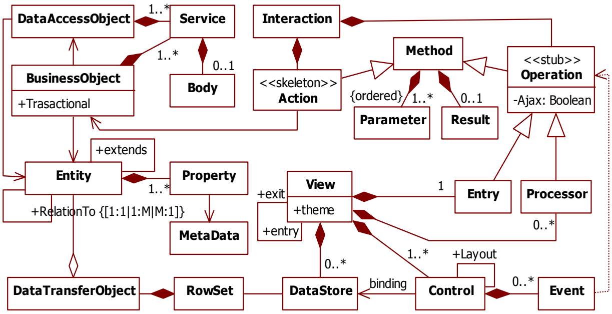


Fig. 2. Simplified metamodel of core WL4EA concepts

- *Data Access Object* (DAO). The model of data persistence objects that create, retrieve, update, and delete (CRUD) business data.
- *Entity*. The model of domain business data which contains the *properties* and *metadata*.

Figure 2 shows the main model elements of WL4EA. Our model elements for an EA conforms to several EA patterns [19], e.g., the DAO and *data transfer object* (DTO) which aggregates several entities to communicate between client and server to reduce the number of net traffic. In the following sub-sections, we will discuss these elements in detail in a data-driven way that begins with data model.

2.3 Data (Entity) Model

The data model stands for the business data to be processed in an application and needs to be persisted in databases. Generally, we call this business data an *entity* which will be mapped to a database table. The syntax for expressing an entity is shown in Listing 1.

Listing 1

```
EntityUnit ::= “Entity” name [“extends” name] “{
                    “Id” name “:” (“GUID”|”Sequence”|”Assigned”)
                     Property +
                    “}”;
Property ::= name “:” (BasicType|Reference);
BasicType ::= (“int”|“string”|“float”|“long”|“date”|“bool”) [“(“
                     (“max”|“min”|“length”) “(” int “)” |
                     (“past”|“future”) “(” date” “)” |
                     “range (” int “;” int “)” |
                     “pattern (” regularExperssion “)” |
                     “not-null” | “email” | “password” “)”];
Reference ::= “<”|“(”|“[” name “>”|“)”|“]”;
int ::=   digit 1-9 [digits];
digit ::=  0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9;
digits ::= digit [digits];
date ::=  <date formatted as YYYY-MM-DD>;
name ::=  <an identifier>;
regularExperssion ::= <a regular expression>;
```

The syntax of data model focuses on properties an entity has, as well as constrains on it and relationship to other entities. We classified five kinds of properties:

- Id is the identity of an entity, which can be generated automatically or manually assigned.
- Common atom property stands for the basic attribute of the entity.
- A one-to-many (Reference) relation to other entities denoted by “<” and “>”.
- A one-to-one relation (Reference) to other entities denoted by “(” and “)”.
- A many-to-one relation (Reference) to other entities denoted by “[” and “]”.

We can declare constraints on the property to conform. An entity will be mapped to a POJO with Hibernate annotations for OR mapping and property validation. For example, a simplified entity **Employee** may be declared as Listing 2. Using term rewrite technique [20], the generate Java code will be shown as Listing 3 (we omit some Hibernate annotations which can be default as the field name (e.g., @columm for field-column mapping).

Listing 2

```
Entity Employee {
    Id :  Assigned
    name : String
    age : int (min(18))
    manager : [Manager]
    projects : <Project>
}
```

However, for a real program generation, there is still some other information to set. For example, the *package* where the entity class be generated to, and the mapping rule from the entity to table (e.g., the table name will be the entity name with a prefix “T_”). In the code above, we use an implicit rule for *foreign key* property mapping (for example, @JoinColumn(name="manager_id") with a suffix “_id” for *manager*).

Listing 3

```
@Entity
public class Employee implements Serializable {
    @Id
    private String Id;
    private String name;
    @Min(value = 18)
    private int age;
    @ManyToOne
    @JoinColumn(name="manager_id")
    private Manger manager;
    @OneToMany(mappedBy="Employee")
    private List <Project> projects;

    public void setId (String id) {this.Id = id;}
    public String getId() {return Id;}
    /*omitting other fields*/
}
```

Usually, such information will be set in a context configuration before the application generation. Another important issue is the abstraction and expressiveness of the data model. There are still many configurations (annotations) to instruct Hibernate for OR mapping, for example, the CRUD policies. We don't model them in our data model to keep it concise because in most cases the default values are OK for these configurations. The tuning operation can be done in some refinements if necessary. Otherwise the data model would be too complex and low level. So, the balance should be considered.

The entity data with metadata (i.e., format and constraints, etc.) are transformed into JSON format to be manipulated by JavaScript in the client. To make the data processed in the client to be propagated in the server side, the data processing traces should be kept. To support this requirement, we designed a model for data history to record the changed data (i.e., new added, deleted, and modified). Fig. 3 illustrates our model, in which we extend the usual POJO model with extra methods to gain the data history information in the UniPojo class, each of which is an element of the RowSet class contained in the DataStore (see next subsection).

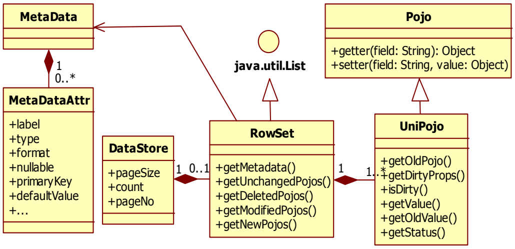


Fig. 3. Data histroy implmentation model

2.4 UI Model

Usually, UI in a WA is *page* centric; such a page in our mode is called a *View*. A view is a quintuple of < *entry*, *Processor*, *DataStore*, *Exit*, *Presentation*>, where

- An *entry* models the accessing point (URL) of a view. A view can have only one entry which can have different arguments for the view.
- *Processor* is a set of CRUD operations over the view on the *DataStore*. Processor = $\{p \mid Status(DataStore) \xrightarrow{p} Status'(DataStore),$ $Status \prec Status' \in \{RA, RD, DM, DL\}\}$, in which, *RA* means "row added"; *RD* means "row deleted"; *DM* means "data modified"; and *DL* means "data reloaded". A processor corresponds to one action in the server side to exchange needed data.
- A *DataStore* is a dataset (a data structure) with metadata information for client computing over the view. Such a dataset can be created from both of the server side and the client side.

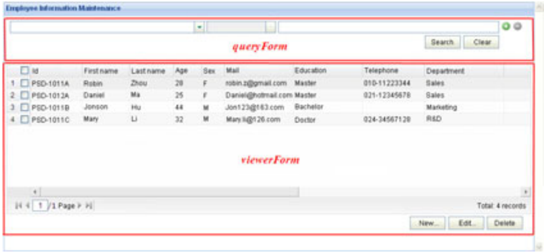
- *Exit* is the set of links for outgoing to other views from the view, i.e., $Exit\ (v) = \{e \mid \exists\ u \in Views, f\,(e) \rightarrow entry \in u \wedge u \neq v\}$, where *v* is the current view, *Views* is the set of views of an application, and *f* is a relation of (*e*, *entry*) that *u* can be reached from *e* via *entry* of *u*.
- *Presentation* is a quadruple of <*style*, *layout*, *Controls*, *Events*>. The elements of *Presentation* are self-explanatory. *Style* is for visualization effects realized by CSS, *layout* is for controls arrangement, and *Controls* are UI widgets which bind to some UI *event* methods.

The simplified syntax for view is shown in Listing 4.

Listing 4

```
View name {
    title : string
    DataStore * name (entity_type ?) {
      columns ? {
          column + : name (label) ?
      }
      parameter * name (type)
      page_size ? : int_number
    }
    entry name ((name : type) *) {
      execution ? string
    }
   processor * name ((name : type) *) {
      onSuccess ? : string
      onError ? : string
   }
   style ? : string (for CSS location)
    container ((border | table | tab | stack | flow)) {
     control * name ((button | checkBox | comboBox
                 | img | dateTextBox | dropDownButton | form
                 | radioButton | grid | tree | label | textArea
                     | menu | fileInput | titlePane | borderPane
                     | contentPanel stackPane | numberTextBox
                     | textbox | password | link)) {
         operationBind * processor_name, control_event
         dataBind ? DataStore_name[[.columns].column]
      }
    }
}
```

The target technologic architecture for view model are JSP (Html), Dojo, and Struts. Therefore, the *title* and *style* represent html header title and CSS declaration, respectively. The *entry* and *processor* represent Ajax methods to be executed before the view is loaded (for *entry*) and interacted with user (for *processor*) via controls' events which are bound to the *processors*. The controls are Dojo UI widgets which are constrained by the *layout* and bind to the *entity* of the DataStore. For example, the view declaration fragment for the window of Fig. 4 is shown in Listing 5.



		Id	First name	Last name	Age	Sex	Mail	Education	Telephone	Department
1	□	PSD-1011A	Robin	Zhou	28	F	robin.z@gmail.com	Master	010-11223344	Sales
2	□	PSD-1012A	Daniel	Ma	25	F	Daniel@hotmail.com	Master	021-12345678	Sales
3	□	PSD-1011B	Jonson	Hu	44	M	Jon123@163.com	Bachelor		Marketing
4	□	PSD-1011C	Mary	Li	32	M	Mary.li@126.com	Doctor	024-34567128	R&D

Fig. 4. A window for employee information mantainance

A typical view will be transformed to six target files, namely, one JSP file for presentation, three JavaScript files for user interaction, and two Java class files for actions control interacting with client end in an Ajax way, thus, an entry or a processor declared in the view model will generate a pair, one is the stub on the client, and the

Listing 5

```
View Employee_Maintenance {
  title : "Employee information maintenance"
  DataStore emp (Employee) { page_size : 10 }
  DataStore query (Employee.Meta)
  processor search (condition : string) {
     …
     queryObject.setQueryObject (condition);
     queryObject.query(emp);
  }
  form (border) {
     queryForm qf {
       operationBind search, qf.b_search.onclick
       dataBind c_condition, query
     }
     viewerForm vf {
       dataBind emp, vf.grid
     }
  }
}
```

other is the skeleton on the sever. For example, the generated code fragment for search processor (for simplicity reason, we use a specific search instead of a general one, which gets all employees of a specific department) will look like as Listing 6.

Listing 6

```
dojo.provide ("Employee.Processor");
dojo.declare ("Employee.Processor", view.Processor, {
   search : function (department,_load,_error) {
      var emp = new DataStore();
      ds.setParameter("department", department);
      return base.Action.requestData ({
         url : "Employee_processor!search.action",
         sync : true,
         load : function(emp) {
            _load&&_load (emp);
            getDateCenter (emp);
         },
         error : function (xhr) {
            _error&&_error (xhr);
         }
      },emp);
   }
});
```

```
@Namespace("/Employee")
public class EmployeeProcessor extends BaseProcessor{
  @Autowired
  private EmployeeBO empBO;
  public void setEmployeeBO (EmployeeBO bo) {
     empBO = bo;
  }
  @Action("search")
  public void search () {
     ViewContext context = generateContext();
     String dept = context.getString("department");
     DataStore ds = DataStoreFactory.getInstance().
     createDataStore("emp");
     List emps = empBO.getEmployeeByDept(dept);
     // transform a List to a DataStore.
     ds = PojoUtil.toDataStore(emps,ds);
     write(ds); // write the DataStore to the client.
  }
}
```

The generated artifacts technical architecture for a view is depicted in Fig. 5. In runtime, the view.jsp (actually, the corresponding html) will create a view JavaScript object (declared in view.js) before it is loaded, during which, the view object will create an entry and a processor JavaScript objects (declared in entry.js and processor.js respectively) in its constructor method. After that, the init() of the view will be invoked to realize some initialization work, e.g., to create some DataStores and invoke the *entry* method of the entry object to get some data from the Entry Java object on the server side.

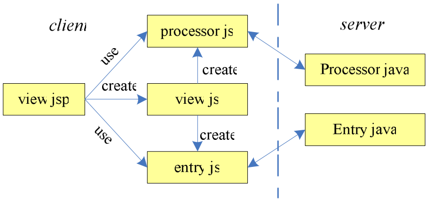


Fig. 5. Generated view (UI) architecture

2.5 Business Logic Model

Business logic refers to the business services or operations for filling the business goals of the application. Since the diversity of the business, a DSL is not competent for it like a general purpose language, such as Java. In WL4EA, we only support some basic business logic patterns like CRUD and general Java methods operations. Thus we can generate default CRUD operations and skeletons of other custom business processing services automatically and refine these business custom services manually. Listing 7 shows the business logic (we call it a BO) declaration specification.

Listing 7

```
BusinessObject name (entity e) {
   operation + ( C | R [Id] | U | D) | name {
   parameter * name : type
   exception * name : type
   return ? returnType
   transactional ? {
      propagation : (required | supports | mandatory
         | requires_new | not_supported | nested |
         never)
      isolation : (serializable | read_uncommited |
         repeatable_read | read_commited | default)
      }
      body ? operationBody
   }
}
```

Listing 8

```
public class EmployeeBO {
   @Autowired
   private EmployeeDAO dao;
   public void setEmployeeDao (EmployeeDAO dao) {
      this.dao = dao;
   }
   @Transactional(propagation=Propagation.required)
   public void save (Employee emp) {
      dao.save (emp)
   }
}

public class EmployeeDAO extends
    HibernateDaoSupport {
   public Serializable save (Employee emp) {
      return getHibernateTemplate().save(emp);
   }
}
```

This specification shows that users can declare CRUD operations on an entity (a special method is R [Id] that indicate a method with the signature of getXXXById (long id) will be generated where XXX represents the actual entity type and we also assume the Id type for such entity is long, otherwise a general method with the signature of getXXX (String str) where the str is a where clause constructed in runtime). All the CRUD method implementations of a BO are proxies of the counterpart methods in a *Data Access Object* (DAO) class with same method signature. The difference between a BO and a DAO is that we can add transactional declaration (the propagation and isolation) in a BO method implemented by corresponding Spring annotations. A DAO is responsible for the actual business data retrieval and persistence in the Hibernate way. Thus, Action (controller), BO, and DAO form a strict layer architecture which assures that an application can be more extensible and maintainable. For example, the generated code fragments of save (Employee emp) in EmployeeBO and EmployeeDAO may look like Listing 8.

Similar to many existing approaches, we don't model the dynamic aspects of an EA, i.e., the body of a business service, since it is still difficult to do and usually written with manual refinements. However, we do support this refinement in the concrete syntax with escape characters, thought it may complicate the DSL programs and make it hard to maintain.

2.6 Implementation

The separated DSL models form a language family that instructs the models composition under the constraints and composition rules of the metamodel. For example, the inclusion of an entity in a view is captured by the *DataStore* construct, and we check whether such an entity type exists or not in a validation activity.

Currently, we use XML as an intermediate model and XML Schema as the metamodel for syntax validation for our DSLs to be transformed to the target code since XML has a well structural semantics and can be easily manipulated by many tools and libraries. Before the final code generation, we first union the separated DSL programs to form a complete one and then check the integrity and coherence of the model elements, and finally, transform these DSL programs to target code (Java, JSP, JavaScript, etc.) by the template-based model transformation [21] using Eclipse JET and XPath. The main role played by JET is transformation control and formatting, while XPath is used to locate needed elements.

3 Related Work

General MDD. In [22], the author proposed a web modeling approach with UML stereotype and tagged value mechanism, and separated a web page into a client one and server one. WebML [23] is a modeling language with ER model and provides a platform-independent conceptual model for data-intensive WA. Autoweb [4] provides code generation mapping between relational data schema and web pages, which targeted for a CGI runtime. The approach in [9] is a MDD approach for WA and targetes for a MVC framework. However, it does not address the interactivity aspect of an application. The authors of [24] proposed a MDD approach for WAs based on UWE method, which is JavaServer Faces (JSF) oriented, while, our approach is JSP oriented, which is dominant in EA development.

DSLs. In [25], the authors proposed a mapping and navigation rule set between data schema and web pages. Such a mechanism can be seen in Rails. HypeDe [26] is an environment underpinned by Rails, which combined SHDM [27] and Ruby to construct a WA rapidly. WebLang [28] is a DSL using Java syntax to encapsulate Java EE elements (servlet, EJB, JSP, etc.) as building blocks. WebDSL [10] is a DSL family for modeling entity, page, access control, and even workflow in dynamic WAs and is divided into a core level and an extended level to make an extension easily. Our approach also supports access control with a security utility implemented in the technical platform performed on a view's entry when it is being called.

4 Conclusion

In this paper, we propose a DSL approach for interactive EAs development. Such a language family includes entity, view, business object, and data access object elements, which form a typical EA. We incorporate an Ajax library to model the Ajax interactions for support high interactivity. We approach targets on lightweight EAs based on some open source application frameworks. Preliminary results showed that our approach can lower the technical complexity and improve the productivity. For future work, we are going to give a parameterization mechanism for our DSL programs to support models reuse. With the advancement of MDD and its synergy with a systematic reuse (e.g., software product line) approach, we hope more mature or even disciplined software development practices in the EA domain.

Acknowledgments. This work was partially supported by the National Significant Science and Technology Projects of China under grant no. 2011ZX01043-001-001.

References

1. Jeff Offutt, Ye Wu.: Modeling presentation layers of web applications for testing. Software and Systems Modeling 9(2): 257–280 (2010)
2. Jeff Offutt.: Quality Attributes of Web Software Applications. IEEE Softw. 19(2): 25–32 (2002)
3. Murugesan, S., Deshpande, Y.: eeting the challenges of Web application development: the web engineering approach. In: ICSE 2002, pp. 687–688. IEEE CS Press (2002)
4. Piero Fraternali, Paolo Paolini.: Model-driven development of Web applications: the AutoWeb system. ACM Trans. Inform. Sys. 18(4): 323–382 (2000)
5. Tommi Mikkonen, Antero Taivalsaari.: Web Applications-Spaghetti Code for the 21st Century. In: SERA 2008, pp. 319–328. IEEE CS Press (2008)

6. Kuuskeri, J., Mikkonen, T.: Partitioning Web Applications between the Server and the Client. In: SAC 2009, pp. 647–652. ACM Press, New York (2009)
7. Mendes, E.: The Need for Empirical Web Engineering: An Introduction. In: Web Engineering: Modelling and Implementing Web Applications, pp. 421–447. Springer, Heidelberg (2007)
8. Pressman, R.S.: Can Internet-Based Applications Be Engineered? IEEE Softw. 15(5), 104–110 (1998)
9. Tai, H., Mitsui, K., Nerome, T., Abe, M., Ono, K., Hori, M.: Model-driven development of large-scale Web applications. IBM J. Research and Development 48(5/6), 797–809 (2004)
10. Visser, E.: WebDSL: A Case Study in Domain-Specific Language Engineering. In: Lämmel, R., Visser, J., Saraiva, J. (eds.) Generative and Transformational Techniques in Software Engineering II. LNCS, vol. 5235, pp. 291–373. Springer, Heidelberg (2008)
11. Frankel, D.S.: Model Driven Architecture™– Applying MDA™ to Enterprise Computing. Wiley Publishing, Inc., Chichester (2003)
12. Schmidt, D.C.: Model-Driven Engineering. Computer 39(2), 25–31 (2006)
13. Mernik, M., Heering, J., Sloane, A.M.: When and How to Develop Domain-Specific Languages. ACM Computing Surveys 37(4), 316–344 (2005)
14. Bell, P.: A Practical High Volume Software Product Line. In: OOPSLA 2007, pp. 994–1003. ACM Press, New York (2007)
15. Kelly, S., Tolvanen, J.-P.: Domain-Specific Modeling: Enabling full code generation. Wiley-IEEE CS Press (2008)
16. Zhang, J., Chung, J.-Y., Chang, C.K.: Towards Increasing Web Application Productivity. In: SAC 2004, pp. 1677–1681. ACM Press, New York (2004)
17. Greenfield, J., Short, K., Cook, S., Kent, S.: Software Factories: Assembling Applications with Patterns, Models, Frameworks, and Tools. Wiley, Chichester (2004)
18. Cuadrado, J.S., Molina, J.G.: A Model-Based Approach to Families of Embedded Domain-Specific Languages. IEEE Trans. Softw. Eng. 35(6), 825–840 (2009)
19. Fowler, M., Rice, D., Foemmel, M., Hieatt, E., Mee, R., Stafford, R.: Patterns of Enterprise Application Architecture. Addison Wesley, Reading (2002)
20. Baader, F., Nipkow, T.: Term Rewriting and All That. Cambridge University Press, Cambridge (1998)
21. Czarnecki, K., Helsen, S.: Feature-based survey of model transformation approaches. IBM Systems Journal 45(3), 621–645 (2006)
22. Conallen, J.: Modeling Web Application Architectures with UML. Commun. ACM 42(10), 63–70 (1999)
23. Ceri, S., Fraternali, P., Bongio, A.: Web Modeling Language (WebML): A Modeling Language for Designing Web Sites. Computer Networks 33(1-6), 137–157 (2000)
24. Kroiss, C., Koch, N., Knapp, A.: UWE4JSF: A Model-Driven Generation Approach for Web Applications. In: Gaedke, M., Grossniklaus, M., Díaz, O. (eds.) ICWE 2009. LNCS, vol. 5648, pp. 493–496. Springer, Heidelberg (2009)
25. Cadavid, J.J., Lopez, D.E., Hincapié, J.A., Quintero, J.B.: A Domain Specific Language to Generate Web Applications, `http://www.archetypus.net/MarTE/doc`
26. Nunes, D.A., Schwabe, D.: Rapid prototyping of web applications combining domain specific languages and model driven design. In: ICWE 2006, pp. 153–160. ACM Press, New York (2006)
27. Lima, F., Schwabe, D.: Modeling Applications for the Semantic Web. In: Cueva Lovelle, J.M., Rodríguez, B.M.G., Gayo, J.E.L., del Pueto Paule Ruiz, M., Aguilar, L.J. (eds.) ICWE 2003. LNCS, vol. 2722, pp. 417–426. Springer, Heidelberg (2003)
28. Buchwalder, O., Petitpierre, C.: WebLang: A Language for Modeling and Implementing Web Applications. In: SEKE 2006, pp. 584–590 (2006)

Scalable Application Description Language to Support IPTV Client Device Independence Based on MPEG-21

Tae-Beom Lim[1,2], Kyoungro Yoon[2,*], Kyung Won Kim[1], Jae Won Moon[1],
Yun Ju Lee[1], and Seok-Pil Lee[1]

[1] Korea Electronics Technology Institute, #68 Yatop-dong,
Bundang-gu, Seongnam, Gyeonggi, 463-816 Korea
[2] Konkuk University, 1 Hwayang-dong, Gwangjin-gu, Seoul
143-701 Korea
tblim@keti.re.kr, yoonk@konkuk.ac.kr,
{kwkim,jwmoon,yjlee0618,lspbio}@keti.re.kr

Abstract. This paper presents a framework and a description language to ensure application interoperability for heterogeneous IPTV client devices with different capabilities by providing application scalability. To support device independent application in IPTV service environment, we suggest a new XML schema (named Scalable Application Description Language: SADL) based on MPEG-21 DIDL (Digital Item Declaration Language).

Keywords: Application Scalability, IPTV, MPEG-21, Digital Item, XML.

1 Introduction

In IPTV service environment, various kinds of applications based on XHTML, JAVA, or Flash with bi-directional interactive functionality are deployed. Many IPTV service providers are targeting to provide their services through various consumer devices such as high-definition TVs with IPTV set-top box, personal computers, mobile phones, etc. For the success of services through different types of devices, suitable multimedia contents for various IPTV client devices should be provided.

In this paper, we propose a new description language to provide scalability of IPTV applications in a form of XML schema called Scalable Application Description Language (SADL). We design SADL to accommodate functionality of selecting and filtering applications or fragments of an application based on the factors such as user's viewing states, or delivery context based on MPEG-21 DIDL [1], [2]. Fig. 1 shows the concept of scalable application framework[4]. In this framework, an application is divided into items and items are further divided into components, which can be considered as fragments of an application. This set of items and components is reconstructed to be optimally presented at each individual environment in the adaptation process, based on the delivery context.

[*] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 361–368, 2011.

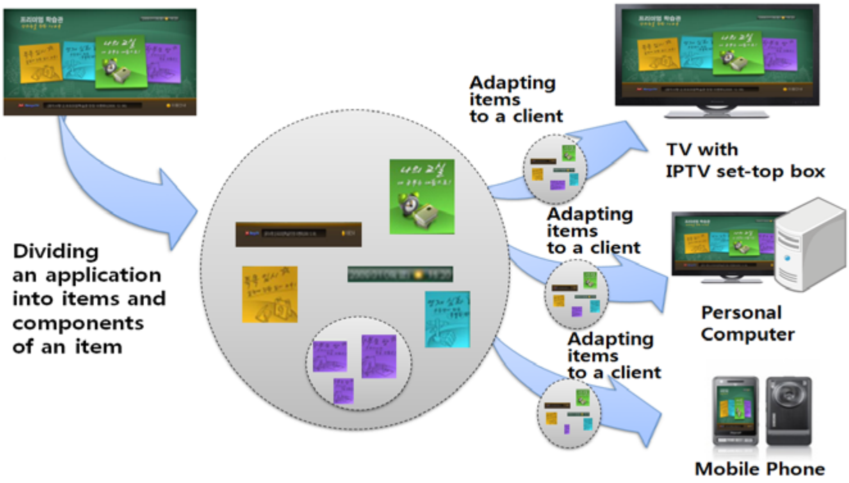

Dividing an application into items and components of an item
Adapting items to a client
TV with IPTV set-top box
Adapting items to a client
Personal Computer
Adapting items to a client
Mobile Phone

Fig. 1. Concept of Scalable Application Framework

2 Scalable Application Description Language

SADL is a new description language in a form of XML schema, and can describe functions of selecting and filtering applications or fragments of an application based on the factors such as device capabilities, user's viewing states, the user profile, or service policies. In addition, we design SADL to be harmonized with MPEG-21 DIDL easily.

2.1 DCConditions

We define a new complex type named *DCCondition* to describe conditions for selecting and filtering Digital Items to present an adapted application. *DCCondition* type is extended from *StatementType* type of DIDL for the harmonization with DIDL. *StatementType* type provides a textual value that contains descriptive, control, revision tracking or identification information [1]. Also, to minimize the number of letters used in a XML document and the depth of the XML tree of the condition description, it is proposed to use the Reverse Polish Notation (RPN), which can be easily implemented using stack-based functions[5], to express complex conditions based on mathematical expressions such as Boolean expressions, comparison expressions or arithmetic expressions. Fig. 2 shows the structure of *DCConditionType* type. *DCConditionType* type can contain *StackEntry* elements to describe condition expressions for selecting and filtering Digital Items as the RPN form[6][7][8].

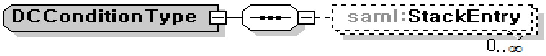


Fig. 2. The structure of *DCConditionType* type

2.2 StackEntryType

StackEntryType type is a base abstract type to represent an operator or an operand for describing stack entries that are members of the stack operation. In SADL, an operand of the stack function may contain a constant value, or a DCFunction that is a function for describing the characteristics of the device, the network, the user profile, and any other aspects that affect execution and presentation of applications on a client device. Fig. 3 shows the classification of the stack entries and how to extend the *StackEntryType* type.

The stack entries are classified into three groups of types based on the data type of the stack entry or the return value type of the stack entry, when the stack entry is used as a stack operator or DCFunction, or. These three new groups of entry types that are extended from *StackEntryType* type are *BooleanEntryType* type, *StringEntryType* type, and *NumericEntryType* type.

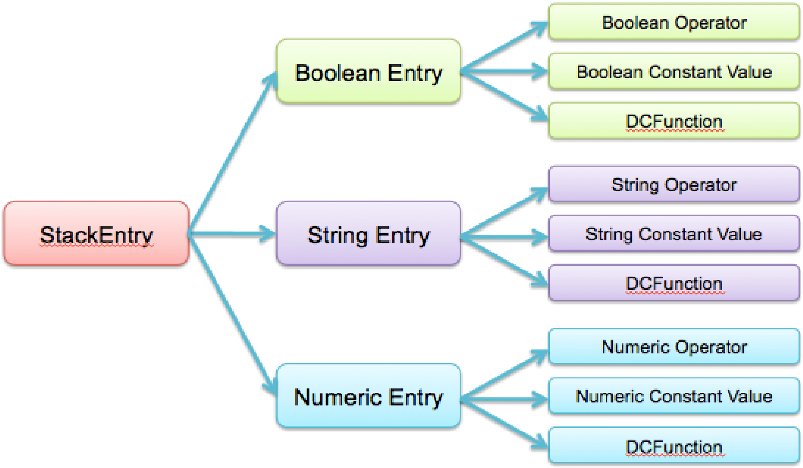


Fig. 3. The classification of the stack entries and the extension rule of the *StackEntryType* type

2.3 Operator Types

The operator types are subclasses of the stack entries, and allow building the stack operation trees. There are three groups of operator types: the Boolean operator group, the string operator group, and the numeric operator group. Each group is composed of the operators that have the same return value type as a result of the operation. In SADL, only the return value type is considered as the criteria of grouping operators, as it is not straightforward to identify the operand data types of a specific operator before actual stack operations are performed due to the characteristics of stack operations.

2.4 DCFunction Types

The DCFunction types define the functions that check and return input factor values that may influence the application adaptation of the SADL filtering engine in the server, the intermediary, or the client. Table 2 shows the characteristics of DCFuction types. It is not an exhaustive list of the DCFunction types, but is a list of minimum set of DCFunction types used in our system.

Table 1. List of operator types

Class	Operator Type	Return Value
Boolean	*AND, OR, NOT, XOR*	Boolean
Comparison	*EQ, NEQ, GT, GTE, LT, LTE*	Boolean
String	*Contains*	Boolean
String	*UpperCase, LowerCase*	String
Numeric	*Add, Subtract, Multiply, Divide, Modulus,Abs, Ceiling, Floor, Round*	Numeric

Table 2. The Characteristic of the DCFunction types

Characteristics	Information
Device Capability	Client Device type, Operation System, Supporting Codec (Video, Audio, Image, etc), Input Interface, Output Interface, Screen Resolution, Virtual Machine, etc
Device State	Power Level, Orientation of Screen, etc
User Information	User Profiles, User Preference, Usage History
Connection	Bandwidth, Network Protocols, Latency, etc
Location	Geographic Coordinates, Time of Day
Locale	Local Language, Local Time Zone
Environment	Noise, Light, Temperature, etc
Service Policy	Subscription Status, Content Restriction, Security, Privacy, etc

Because there are many aspects of the input factors, we categorize the DCFunction types for the easy extensibility and manageability of the SADL schema. Fig. 4 is the diagram of categories of the DCFunction types of four levels. In the first level, we divide the DCFunction types into three categories of *Client, Intermediary* and *Sever*, according to the device expected to process the function described in the DCFunction type. In the second level, each category of the DCFunctions is subdivided into two subcategories of *Static Return Value* and *Dynamic Return Value* based on the flexibility of the return value of function. The DCFunction is categorized as *Dynamic Return Value* if the return values can change dynamically. Otherwise, it is categorized as *Static Return Value*. In the third level, each of the categories at the second level is categorized further into four subcategories of *Device Capabilities/State, User Information, Viewing Environment/State,* and *Service Policy* based on the characteristics of the functions. The *Device Capability/State* category includes the information concerned with the device capabilities, the device state and the connection; the *Viewing Environment/State* category includes the information of the location, the locale, and the environment. In the last level, the DCFunction types are categorized into three categories of Boolean DCFunctions, String DCFunctions, Numeric DCFunctions, based on the type of the return value, in the same way as the constant value type, and the operator type.

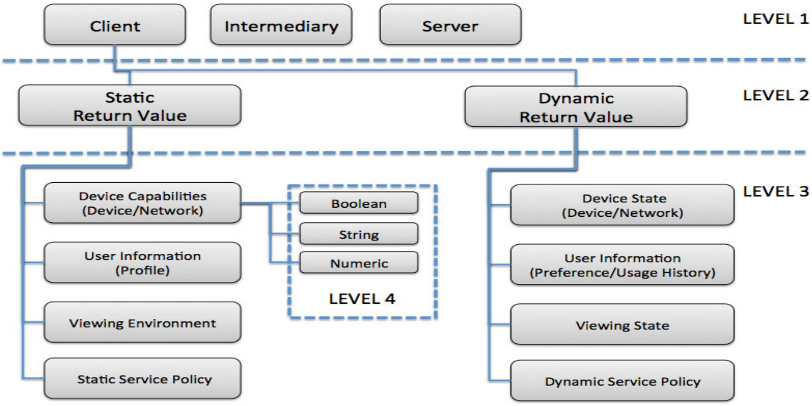


Fig. 4. The categorization of the DCFunction type

The DCFunctions are classified into sub-types using the typing mechanism of the XML schema to enable the validation mechanism of the SADL parser for the validity check of the function types in the process of parsing a SADL instance document.

Boolean DCFunction Types. The Boolean DCFunction types are the functions that should return a Boolean value of TURE or FALSE. These types are extended from the *BooleanEntryType* type, which is a base abstract type. Fig. 5 shows a snippet of the Boolean DCFunction types in SADL XML schema. Fig. 6 shows a snippet of SADL instance document, in which the condition given in the following example is described.

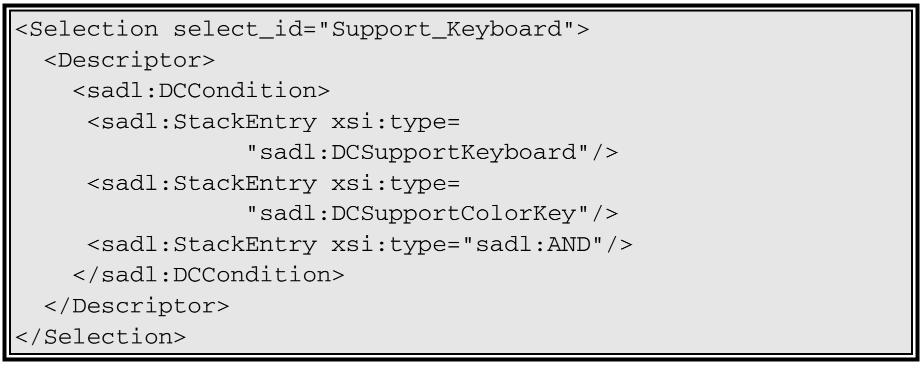
((Has the device a keyboard interface?) &
(Are there color keys in keyboard?))

```
<complexType name=" DCSupportKeyboard">
        <complexContent>
                <extension base="sadl:BooleanEntryType"/>
        </complexContent>
</complexType>

. . .

<complexType name=" DCSupportHTTP">
        <complexContent>
                <extension base="sadl:BooleanEntryType "/>
        </complexContent>
</complexType>
```

Fig. 5. Snippet of the Boolean DCFunction types



```
<Selection select_id="Support_Keyboard">
  <Descriptor>
    <sadl:DCCondition>
     <sadl:StackEntry xsi:type=
                "sadl:DCSupportKeyboard"/>
     <sadl:StackEntry xsi:type=
                "sadl:DCSupportColorKey"/>
     <sadl:StackEntry xsi:type="sadl:AND"/>
    </sadl:DCCondition>
  </Descriptor>
</Selection>
```

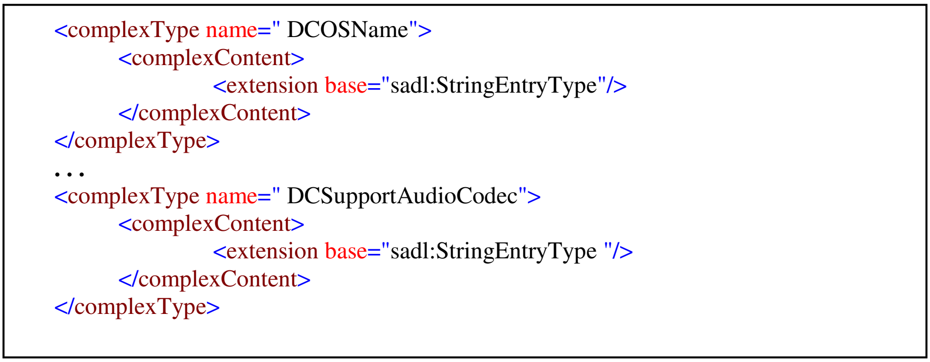
Fig. 6. An example of using the Boolean DCFunction type

String DCFunction Types. The String DCFunction types describe the functions that should return a String value. These types are extended from the *StringEntryType* type, which is a base abstract type. Fig. 7 shows a snippet of the string DCFunction types in SADL XML schema. Fig. 8 shows a snippet of SADL instance document, in which the condition given in the following example is described.

(Does the device support the MP3 audio codec?)

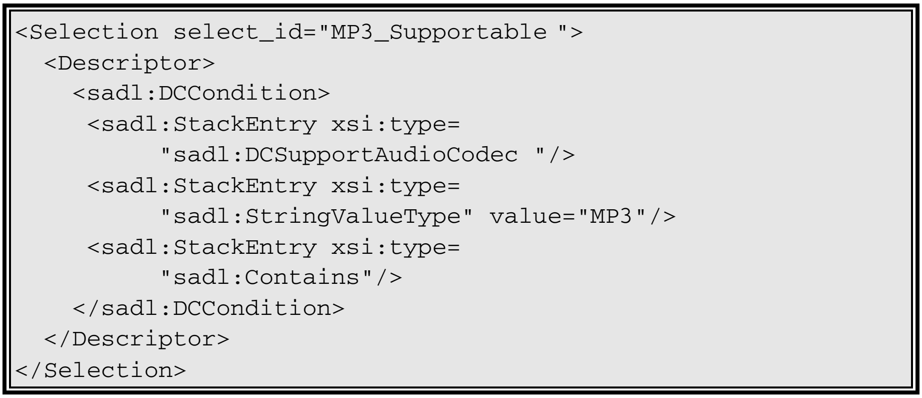
Numeric DCFunction Types. The numeric DCFunction types describe the functions that should return a numeric value. These types are extended from the *NumericEntryType* type, which is a base abstract type. Fig. 9 shows a snippet of the Numeric DCFunction types in SADL XML schema. Fig. 10 shows how to describe the following example condition in a SADL instance document.

((Available memory size of the device) > 1024000)



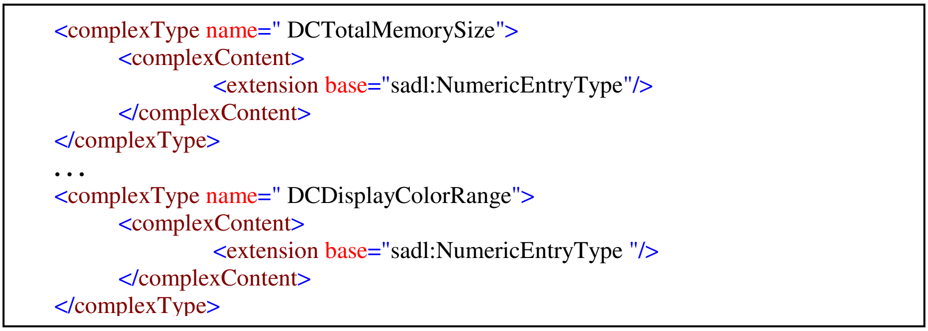
```
<complexType name=" DCOSName">
        <complexContent>
                <extension base="sadl:StringEntryType"/>
        </complexContent>
</complexType>
. . .
<complexType name=" DCSupportAudioCodec">
        <complexContent>
                <extension base="sadl:StringEntryType "/>
        </complexContent>
</complexType>
```

Fig. 7. Snippet of the String DCFunction types



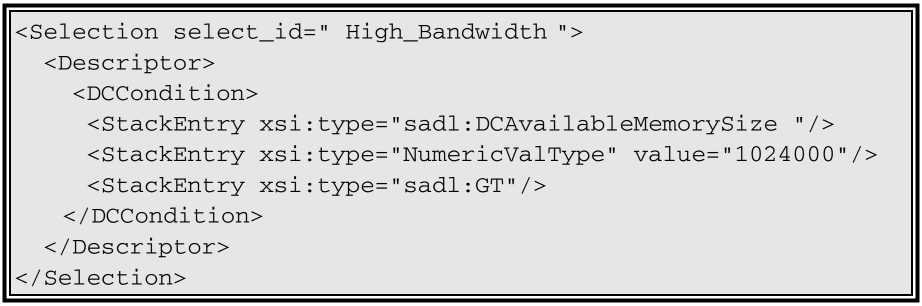
```
<Selection select_id="MP3_Supportable ">
  <Descriptor>
    <sadl:DCCondition>
     <sadl:StackEntry xsi:type=
          "sadl:DCSupportAudioCodec "/>
     <sadl:StackEntry xsi:type=
          "sadl:StringValueType" value="MP3"/>
     <sadl:StackEntry xsi:type=
          "sadl:Contains"/>
    </sadl:DCCondition>
  </Descriptor>
</Selection>
```

Fig. 8. An example of using the String DCFunction type



```
<complexType name=" DCTotalMemorySize">
    <complexContent>
        <extension base="sadl:NumericEntryType"/>
    </complexContent>
</complexType>
...
<complexType name=" DCDisplayColorRange">
    <complexContent>
        <extension base="sadl:NumericEntryType "/>
    </complexContent>
</complexType>
```

Fig. 9. Snippet of the numeric DCFunction types



```
<Selection select_id=" High_Bandwidth ">
  <Descriptor>
    <DCCondition>
     <StackEntry xsi:type="sadl:DCAvailableMemorySize "/>
     <StackEntry xsi:type="NumericValType" value="1024000"/>
     <StackEntry xsi:type="sadl:GT"/>
   </DCCondition>
  </Descriptor>
</Selection>
```

Fig. 10. An example of using the numeric DCFunction type

5 Conclusion

In this paper, we propose a new framework to provide scalability of IPTV applications in a form of XML schema called Scalable Application Description Language (SADL). We design SADL to accommodate functions of selecting and filtering applications or fragments of an application based on MPEG-21 DIDL.

Currently, we are implementing the SADL authoring tool, a SADL delivery server and SADL client parser modules. Also, we are trying to deploy the proposed technologies on a Korean commercial fixed and mobile IPTV convergence service in the near future.

References

1. ISO/IEC IS 21000-2 Information technology – Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration, ISO Publication (October 2005)
2. Burnett, I.S., Davis, S.J., Drury, G.M.: MPEG-21 digital item declaration and Identification-principles and compression. IEEE Trans. Multimedia 7(3), 400–407 (2005)
3. ISO/IEC IS 21000-7 Information technology – Multimedia framework(MPEG-21) – Part 7: Digital Item Adaptation, ISO Publication (December 2007)
4. Lim, T.-B., Kim, K.W., Lee, Y.J., Moon, J.W., Yoon, K.: Scalable application framework to support IPTV client device independence based on MPEG-21. In: 2011 IEEE International Conference Consumer Electronics (ICCE), pp. 856–860 (2011)
5. Barton, R.S.: A New Approach to the Functional Design of a Digital Computer, afips. In: 1961 Proceedings of the Western Joint Computer Conference, pp. 393–396 (1961)
6. XML Schema Part 0:Primer, `http://www.w3.org/TR/2001/REC-xmlschema-0-20010502`
7. XML Schema Part 1:Structures, `http://www.w3.org/TR/2001/REC-xmlschema-1-20010502`
8. XML Schema Part 2:Datatypes, `http://www.w3.org/TR/2001/REC-xmlschema-2-20010502`

A Study on Using Two-Phase Conditional Random Fields for Query Interface Segmentation

Yongquan Dong, Xiangjun Zhao, and Gongjie Zhang

School of Computer Science and Technology, Xuzhou Normal University, Xuzhou, China
tomdyq@gmail.com, xjzhao@xznu.edu.cn, zhanggongjie@126.com

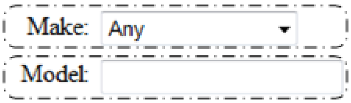
Abstract. Recently, the Web has been rapidly "deepened" by many searchable databases online, where data are hidden behind query interfaces. Automatic processing of a query interface is a must to access the invisible contents of deep Web. This entails automatic segmentation, i.e., the task of grouping related components of an interface together. The segmentation is divided into two steps: interface component labeling and interface component grouping. In this paper we present a new approach to perform query interface segmentation using two-phase Conditional Random Fields (CRFs). At the first phase, one CRFs model is used to tag each component with a semantic label (attribute-name, operator, operand or other); at the second phase, another CRFs model is used to create groups of related components. Experiments show that our approach yields high accuracy.

Keywords: Query Interface Segmentation, Two-Phase Conditional Random Fields, Deep Web.

1 Introduction

In recent years, the problem of retrieving and integrating information available in online databases has received a lot of attentions in the research and industrial[1] communities[1][2][3],because of the quality of the information and the growing number of online databases – it is estimated that there are several million online databases[4].

Since most online databases can only be accessed by filling up query interfaces, to automatically integrate them and retrieve their contents, a prerequisite is an automatic understanding of query interface semantics. This entails automatic segmentation, i.e., the task of grouping related components of a query interface together. For example, there is one query interface in Fig.1 with two segmentations. The upper segmentation contains two components: "Make:", selection list. The bottom segmentation also contains two components: "Model:", textbox.



Make: Any

Model:

Fig. 1. Segmented Query Interface

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 369–376, 2011.

Previous solutions [5][6][7][8] for segmentation employed a variety of techniques including rules, heuristics, and machine learning. However, none of the solutions has explored the Conditional Random Fields (CRFs) [9] model for labeling and grouping interface components. CRFs are discriminatively-trained undirected graphical models that have great freedom to use complex, overlapping and non-independent feature sets. We conduct a thorough study on the effectiveness of using CRFs for interface segmentation.

As the interface segmentation is divided two steps, i.e. interface component labeling and interface component grouping. And the two steps are finished in two phases. At the first phase, all components are tagged by a CRFs model with semantic labels (attribute-name, operator, or operand); at the second phase, another CRFs model is used to create groups of related components. Experimental results using a large number of real-world data collected from diverse domains indicate that the proposed approach yields high accuracy.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the model of Conditional Random Fields and query interface analysis is presented in Section 4. Section 5 presents the approach of query interface segmentation using two-phase CRFs. Section 6 reports the experimental results on the datasets from multiple domains. Section 7 concludes the paper and discusses several directions for future research.

2 Related Works

In this section we describe the existing approaches for query interface segmentation.

The techniques of query interface segmentation can be divided into two categories: rule/heuristic-based and model-based. A large majority of existing works is based on rules and heuristics. Both rules and heuristics are crafted manually by observing the patterns on interface. Model-based approaches adopt a holistic approach and automatically learn the patterns into a model using training data. This learned model is then used for segmenting query interfaces.

An example of a rule-based technique is the work by Zhang et al.[6]. They assume that a hidden syntax guides the presentation of segments (query conditions) on a query interface. A query condition consists of the 3-tuple, {attribute, operator, values}. The tuples are identified using pres-specified grammar rules. An example of heuristic-based techniques is the work by He et al.[5].They proposed a LEX approach, in which the segment, logical attribute, is identified using a set of heuristics. Every HTML form element of an interface is associated with a surrounding text, and preference is given to the text that ends with a colon, matches with element's internal name, or is closer to the element. The form elements associated with the same text and the text itself belong to a logical attribute.

With the growing diversity in interface design [10], it would be difficult for rule-based and heuristic-based techniques to cope with large interface collections. For every new interface, a new set of extraction rules would be required. Nguyen et al.[7] proposed the LabelEx approach which is a path-breaking model-based work based on supervised machine learning. It uses Naive Bayes and Decision Trees classifiers to assign text labels to form elements based on textual and layout features of components.

Ritu Khare et al.[8] also proposed a model-based technique HMM-IS which explored Hidden Markov Models to create an artificial designer that has the ability to segment an interface. It complements the LabelEx approach, in that along with label assignment, it also groups the related components together.

In this paper, we explore another model-based technique, Conditional Random Fields. We use the two-phase CRFs to segment query interfaces. The proposed approach improves the process of manual rule-based and heuristic-based techniques in that it automatically learns the rules and heuristics into the model. It complements the model-based approach LabelEx, in that along with label assignment, it also groups the related components together. It also performs better than HMM-IS, because unlike HMM, CRFs support the user of many rich and overlapping features.

3 Conditional Random Fields

A conditional random field is an undirected graphical model that defines a single exponential distribution over label sequences given a particular observation sequence. That is to say, Let $G=(V,E)$ be an undirected graph, where V is the set of states Y={y$_i$|1≤i≤n} for a given length-n input sequence X=x$_1$…x$_n$ and E={(y$_{i-1}$,y$_i$)| 1≤i≤n } is the set of n-1 edges in the linear chain. Following Lafferty et al. [9], the conditional probability of the state sequence Y for a given input sequence X is

$$p_\Lambda(y\mid x)=\frac{1}{Z(x)}\exp\left\{\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,x,t)\right\} \tag{1}$$

where $f_k(y_{t-1},y_t,x,t)$ is a feature function of vertex and edge, λ_k is a learned weight associated with f_k, Λ is the weight set, $\Lambda=\{\lambda_k\}$, and $Z(x)$ is the normalization factor, also know as partition function, which has the form,

$$Z(x)=\sum_{y}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(y_{t-1},y_t,x,t)\right) \tag{2}$$

Training of CRFs requires estimating the values of the weight set, Λ , which is usually done by maximizing the log-likelihood of a given training set. Popular training methods include generalized iterative scaling, conjugate-gradient and limited-memory quasi-Newton [11].

Once these weights are found, the labeling for a new unlabeled sequence can be done using a modified Viterbi algorithm.

4 Query Interface Analysis

A query interface consists of a sequence of components that belong to different logical groups, i.e., segment. Components in a single segment have difference semantic roles. For example, in Figure 2, each component with the dotted rectangle is labeled with a text, which we term as a semantic label. For a given component, the associated semantic label denotes the meaning of the component from user's or designer's standpoint. Web designer do not usually embed explicit component labels in the HTML source code. This makes automatic semantic labeling a difficult task.

In this paper, we use a fixed set of 4 semantic labels, {attribute-name, operator, operand, other} to tag interface components. For deep Web applications, the user-entered values and the text-labels of a query interface are often translated into structured query expressions against the underlying databases. A typical structured query statement, i.e., SQL, contains 3 types of clauses: a clause indicating output(SELECT), a clause pointing to the database(FROM), and a clause specifying query conditions(WHERE). For example, assuming the underlying database table name is "House", following SQL queries might be generated for the interface in Fig. 2:

1. SELECT * FROM House WHERE Realtor= "Black Bear Realty";
2. SELECT * FROM House WHERE Square_footage>=30;
3. SELECT * FROM House WHERE Price_range>=100.

For a typical Web application, although the underlying database name and schema are invisible, some clauses in a structured query are observable. In particular, a response page presenting query results corresponds to a SELECT clause, and a query interface collecting values for query conditions corresponds to a WHERE clause.

A WHERE clause consists of a set of predicates, e.g., Realtor= "Black Bear Realty". Such a predicate often specifies a query condition, using a built-in operator, for a particular attribute in the underlying database schema. Based on this observation, we use semantic labels, attribute-name, operator, operand, and other, for tagging interface components. And the semantic label "other" is used to tag the components not belonging to any segment.

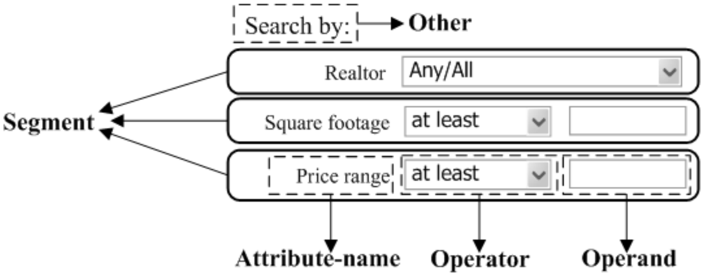


Fig. 2. Assigning Semantic Labels to Components

5 Query Interface Segmentation Using Two-phase CRFs

To explore effective methods for query interface segmentation, this paper presents a two-phase CRFs approach for this task.

5.1 First Phase of Our Two-Phase Approach

At the first phase of our two-phase approach, semantic labels are used to tag the interface components by a CRFs model. In practice, we regard each component in a interface as a token and each token is tagged with a semantic label.

Features used at this phase are shown as follows:

(1)Binary Feature

Binary feature represents if each interface component is the following type: Text, Textbox, Textarea, File Input, CheckBox group, RadioButton group, Selection list.

(2)Text Feature

If the interface component is a Text, it should be described by text features. Otherwise, all text features of the interface component are zero. Text features are shown in Table 1.

Table 1. Text Features

All words start with capitalized letter
Initial word starts with capitalized letter
Number of digits in the text
Number of words in the text
Contain the punctuation
Contain the parentheses
Contain the colon
Contain special keywords “between”
Contain special keywords “min”,”lower limit”,”from”
Contain special keywords “max”,”to”,”less than or equal to”

5.2 Second Phase of our Two-Phase Approach

At the second phase, another CRFs model is used for grouping interface components. We used BIO model [12], where B represents that the interface component is the start of segment, I stands for the inside of segment, and O indicates that the interface does not belong to any segment.

Features used at this phase are shown as follows:

(1) Binary Feature

Binary feature represents if each interface component is the following type: attribute-name, operator, operand, other.

(2) Context Feature

To model local context simply, neighboring interface component in the window [-1,1] are also added as features. For instance, the middle component in the sequence “operator” would have features Current=operator, Neighbor (-1) =attribute-name, Neighbor (+1) =operand.

6 Experiments and Results

6.1 Datasets

The experiments were carried out on 5 representative domains: airline, automobiles, books, jobs and real estate. The details of datasets are shown in Table 2. The first column of Table 2 is the domain of datasets. The second column is the data sources from which we collect the interfaces. The third column is the size of the interfaces. In the last line, we give the urls of the data sources.

Table 2. Sources of Datasets

Domain	Data Source	Size
airline	ICQ, Tel-8, Completeplanet	100
automobiles	ICQ, Tel-8, Completeplanet	100
books	ICQ, Tel-8, Completeplanet	100
jobs	ICQ, Tel-8, Completeplanet	100
real estate	ICQ, Completeplanet	100
ICQ: http://metaquerier.cs.uiuc.edu/repository/datasets/icq/index.html Tel-8 : http://metaquerier.cs.uiuc.edu/repository/datasets/tel-8/index.html Completeplanet : http://www.completeplanet.com/		

The query interfaces were trained and tested using 10-fold cross-validation method. The training sets were first tagged by hand. In both phrases, training and testing were performed using Maximum Likehood and Viterbi algorithms, respectively.

6.2 Discussion on Experimental Results

We have designed two experiments to evaluate the effects of our approach.

6.2.1 The Result of Two Phrases

We measured the results at two phrases of the process. The first phrase is determining the accuracy of interface component labeling and the second phrase is determining the accuracy of interface component grouping. We select airline domain to do the experiments because each interface in this domain has many interface elements. Table 3 shows the results obtained in the empirical evaluation. In the first phrase, we only calculated the accuracy only for the segments which are the whole correctly identified. Overall, 95.67% of attributes-names were correctly identified which is shown in row 1. The tagging accuracies for the operators and operands are shown in rows 2 and 3, respectively. Overall, 91.25% of the operators were correctly identified. A misidentification occurred because an operator was mistaken as one of the other 3 semantic labels. Overall, 98.68% of the operands were correctly identified. In all the incorrect cases, the operand was misidentified as an operator. In the second phrase, the segmentation accuracy is listed in the last row of Table 3. Out of the total of segments present in all the interfaces, 93.13% were correctly identified. A misidentification occurred due to one or more misidentified semantic labels in the first phrase. From the experimental results, we can see that our approach can achieve the high accuracy.

6.2.2 Compared with the Other Approach

We measure the segmentation accuracy achieved on our approach using two-phase CRFs. We compare the approach to a previous heuristic-based solution, LEX [5]and a model-base solution, HMM-IS[8].

Table 3. Segmentation and Tagging Accuracy

Phrase	Semantic Label	Accuracy (%)
Phrase 1	attribute-name	95.67
	operator	91.25
	operand	98.68
	other	80.12
Phrase 2	segmentation	93.13

We implemented the LEX algorithm and HMM-IS algorithm based on the description in [5]and [8]. Table 4 shows the results obtained in the comparable experiments. The first column of Table 4 shows the different domain used in the experiments. The second and third columns show the accuracies attained by LEX and HMM-IS on interfaces from 5 domains, respectively. The fourth column shows the accuracy of our approach.

Table 4. Comparison of Segmentation Accuracy

Domain	LEX(%)	HMM-IS(%)	Our Approach(%)
airline	61.23	78.57	93.13
automobiles	56.58	80.06	88.67
books	63.27	77.56	86.87
jobs	70.78	75.48	90.15
real estate	66.58	79.06	89.08

The result shows that our approach improves accuracies by 19.37%-32.09% over LEX. The main reason is that LEX does not model miscellaneous texts such as "e.g., "10.0-40.0",'. It thus suffered from under-segmentation in a large number of cases. And our approach also improves accuracies by 8.61%-14.67% over HMM-IS. This is because of the conditional nature of CRFs, which results in the relaxation of the independence assumptions required by HMMs.

7 Conclusions

In this paper, we introduced a new approach to perform query interface segmentation using two-phase CRFs. We carried out the segmentation in two phrases: interface component labeling and interface component grouping. At the first phase, we try to tag each component with a semantic label (attribute-name, operator, operand, or other); at the second phase, another CRFs model is used to create groups of related components. We have implemented a prototype and tested it using a large dataset that contains real-world query interfaces in five different domains. The experimental results demonstrate the feasibility and effectiveness of our approach.

Acknowledgement. The work is supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant No. 11KJB520019.

References

1. Wu, W., Yu, C., Doan, A.H., Meng, W.: An interactive clustering-based approach to integrating source query interfaces on the deep Web. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 95–106 (2004)
2. Dong, Y., Li, Q., Ding, Y., Peng, Z.: ETTA-IM:A deep web query interface matching approach based on evidence theory and task assignment. Expert Systems with Applications 38(8), 10218–10228 (2011)
3. Chang, K.C., He, B., Zhang, Z.: Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web. In: Conference on Innovative Data Systems Research, pp. 44–55 (2005)
4. Jeffery, S.R., Cohen, S., Dong, X., Ko, D., Yu, C., Halevy, A.: Web-scale Data Integration: You can only afford to Pay As You Go. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 342–350 (2007)
5. He, H., Meng, W., Lu, Y., Yu, C., Wu, Z.: Towards Deeper Understanding of the Search Interfaces of the Deep Web. World Wide Web 10(2), 133–155 (2007)
6. Zhang, Z., He, B., Chuan, K.C.: Understanding Web query interfaces: best-effort parsing with hidden syntax. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 107–118 (2004)
7. Nguyen, H., Nguyen, T., Freire, J.: Learning to extract form labels. Proc. VLDB Endow. 1(1), 684–694 (2008)
8. Khare, R., An, Y.: An empirical study on using hidden markov model for search interface segmentation. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management, pp. 17–26 (2009)
9. Lafferty, J.D., Callum, A.M., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289 (2001)
10. He, B., Patel, M., Zhang, Z., Chang, K.C.: Accessing the deep web:A Survey. Communications of the ACM 50(5), 94–101 (2007)
11. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. 45(3), 503–528 (1989)
12. Yang, Z., Lin, H., Li, Y.: Exploiting the contextual cues for bio-entity name recognition in biomedical literature. J. of Biomedical Informatics 41(4), 580–587 (2008)

Key Techniques Research on Water Resources Scientific Data Sharing Platform

Yufeng Yu[1], Shijin Li[1,2], and Jingjin Jiang[1,3]

[1] College of Computer & Information, Hohai University, Nanjing 210098, China
[2] National Engineering Research Center of Water Resources Efficient Utilization Engineering Safety, Nanjing 210098, China
[3] College of Jin Cheng, Nanjing University of Aeronautics & Astronautics, Nanjing 211156, China
yfyu@hhu.edu.cn

Abstract. In order to integrate distributed and isomerous water resources scientific data and provide one-stop data sharing services for different users, a distributed data sharing platform is needed urgently in China. A framework of the distributed water resources scientific data sharing platform (WSDSP) is proposed to solve this problem .This platform comprises one main center, one authentication center and several sub-centers in which core functions are encapsulated into web services. And then, some key techniques, i.e. metadata directory service, single sing-on (SSO), web services based service-oriented architecture (SOA) and distributed metadata synchronization, are discussed in detail. And now, this platform was deployed in Ministry of water resources and related provinces, in which more than 8976 MB data integrated from 5487 hydrometric stations can be shared to the public.

Keywords: Data Sharing, Metadata SOA, Single Sing-on, Web Services.

1 Introduction

With increased complexity of scientific problems, scientific research is increasingly a collaborative effort across multiple institutions and disciplines. Scientific researchers need an effective infrastructure to share their original scientific data resources.

As an important part of scientific data resources, water resources data include water resources, flood and drought, water saving irrigation, water conservation, hydrology, hydraulic engineering; which act on building of national economy construction and macroscopic decision-making[1].

However, it is noted that such data resources are scattered in different sectors or organizations and so it is very difficult to share them for scientific research. It may often encounter a paradox that many organizations have accumulated more and more data resources which not used usually while some scientists often complain that there are no available data resources to use. In order to share scientific data to improve their value and avoid investing money repeatedly, the water resources sciences data sharing platform (WSDSP), one of National Infrastructure and Facility Development Environment Building for Science and Technology Industries Program of China, was

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 377–384, 2011.

launched in 2003 sponsored by the Ministry of Science and Technology of China. Its main objective is to integrate scattered and isomerous water sciences data and share them among all scientists to support scientific research.

With several years' practices of this program, this paper introduces the design and implement of the WSDSP as well as some core techniques used in it.

2 The Status of WSDSP

Water resources scientific data is one of the important parts which constitute the fundamental scientific data resources of the nation. It plays an irreplaceable role in the field of floods and droughts control, water resources utilization and protection, science researches and technology activities.

However, for a long time the service ability of data sharing is low and the work of data sharing doesn't have substantial progress for lacking of sharing standards, regulations and mechanism. Especially in the scientific research and education field, these data belongs to different organizations and in fact most data holders do not like to share their data initiatively if their intellectual property rights cannot be protected [2]. Considering the above conditions, we put forward a valid mechanism to manage metadata centrally and store data dispersedly. In addition, for adapting the requirement of management and sharing of different data, it deploys many data sharing web sites according to the data characterization of different disciplines and regions. We customize the data submission, management and presentation services in each web site by which it can easily integrate data of special discipline and region and ensure its quality.

For data users, however, they just care about how to search and get their data one-stop. So it is required that the above data sharing web sites should have the interoperable capacity to provide one stop data sharing service for users.

3 System Structure of WSDSP

The aim of WSDSP is to form a "geographically distributed and logically integrated" distributed sharing platform which can provide inter-discipline and cross-department data sharing services. Therefore, the WSDSP adopts "main center - sub center - data resource" structure mode.

As it shown in fig.1, the WSDSP can be divided into five levels, i.e. portal layer, web service layer, business logic layer, data resource layer and running environment layer. Registered user can obtain data query, browse, download and submit services through portal. Web services layer is the basis of interoperability by realizing several services. Business logic layer controls and schedules specific business logic through the transparent access mechanism of metadata. Data resource layer is the core of WSDSP; it covers some distributed database and manages data resource such as integration, collection and release through the background processing. And running environment layer provides technical support such as database operation and system service.

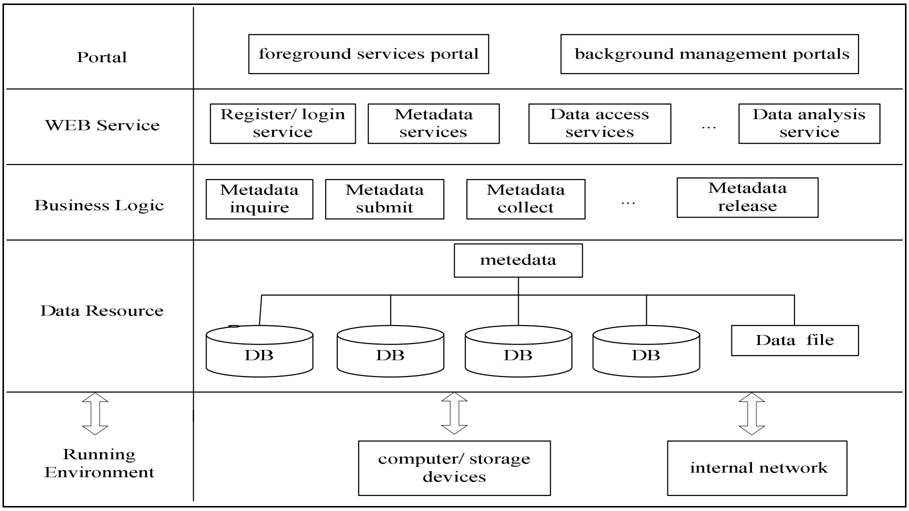


Fig. 1. WSDSP System Structure

The metadata is the foundation of the one-stop service and the transparent layer of the whole WSDSP, which provides data discovery, data query, data description, data navigation services for data users as well as metadata submission, upload and registration services for data producers of distributed data resource.

4 Detailed Designs of Key Techniques

In this section, we discuss some key techniques applying in the WSDSP, including Metadata Directory Service(MDS), Single Sing-On, Web Services based SOA, Distributed Metadata Synchronization and so on. Due to space limitation, we put emphasis on the following key techniques.

4.1 Metadata Directory Service

MDS is the basis of data sharing which presents data resources to the user in the form of dynamic cataloging. The architecture of MDS is shown in figure 2. Data providers generated metadata document and submitted it to sharing platform by using metadata standards' description; Data managers censored metadata information and published it to the metadata register center; Data users accessed or applied data resources through metadata description information in metadata document getting from the metadata registry center.

MDS services for data providers, data managers and data users to submit, censor, publish and query the metadata [3]. It can be established through the following procedures:

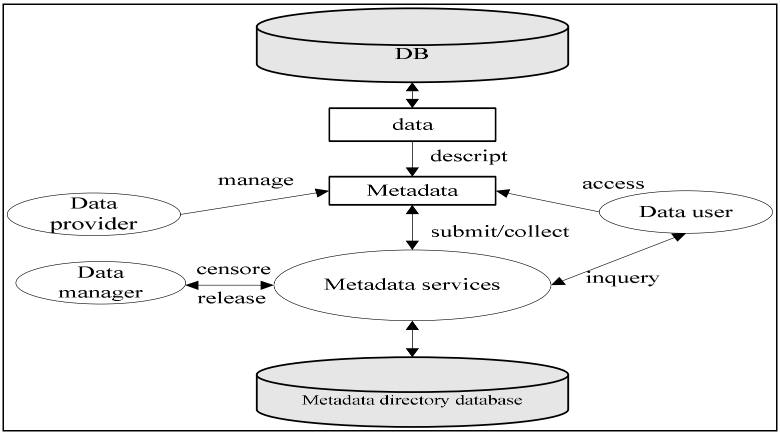


Fig. 2. Metadata directory structure

1) Establish metadata directory database. Data providers submit metadata document to data sharing platform, data managers publish it to the metadata directory service and save it to metadata database after data efficacy and safety censored. MDS provides query, access metadata document services through space database engine and database management system.

2) Indexing. Once a metadata document is published to the MDS system, it requires creating index of the metadata document information for querying and retrieving conveniently.

3) Assign permissions. It can define three different roles according to different users' demand: MetaBrowser, MetaPublisher and MetaAdmin. The users of MetaBrowser role have the authority to search and browse the metadata documents of MDS. A MetaPublisher-role user can search, browse and publish the metadata documents, while a MetaAdmin-role user can do all things of MetaPublisher-role, furthermore, it can update the metadata documents.

4) Provide metadata publish and access interface. Data providers use publish interface to distribute the metadata document generated by metadata standard description; while data users can search, browse metadata through access interface of metadata services.

4.2 Single Sing-On

As Jan de Clercq [10] defines, SSO has the ability for a user to authenticate once to a single authentication authority and then access other protected resources without re-authentication. By adopting SSO system, users can access subsystem and resource in an application by logging once.

In WSDSP, we propose a simple SSO solution adopting central authentication architecture based on HTTP redirection and tickets, with cross-domain cookie sharing as key technology. We deploy a Certificate & Authentication (CA) in the main center to provide uniform user permission management and authentication, meanwhile, we deploy CA client in every sub-center, which can join authentication system by assigning the address of CA. The SSO and resource access control/ authentication can be implemented through the CA deployed in the main-center and CA client deployed in the sub-center [4].

There will be many data centers in the distributed shared service system. Furthermore, the number of data centers would grow with the passage of time. So, the data center doesn't know whether the user has logged in other centers when a user visits it, and it's impossible to traverse each node. Thus we store the users' logon TokenID and access information in the Cookie of authentication center to enable different clients to find the TokenID with a fixed method to judge whether the client has logged in the whole sharing network system. The basic steps are as follows:

1) The user makes his register and gains his authority in CA center.

2) Once a client visits some data resources from a data center, the SSO module takes over the client requests to inquire whether it holds this data center's Cookie and try to get the TokenID stored in the Cookie. If the client can provide, it means that the user had visited this website with legal identity and didn't logged out locally, go to Step 5, otherwise, follow the next steps.

3) CA redirects the client browser to login authentication center and tries to obtain the Cookie stored in the client. And if the Cookie doesn't exist, it means that the user has not been a global registry and go to Step 4;otherwise get TokenID information from CA.Then CA redirects client browser back to the originally access data center. The data center obtains TokenID and return-address, writes the TokenID and user authority information into the client Cookie and jumps to the return-address, then back to Step 2.

4) Redirect the client browser to the login page and confirm user identity by checking login name and password with user information table on the CA. If authentication is successful, generate a unique TokenID randomly and insert an entry into the certificate database table for this user's login. And then save TokenID and user authority information into the cookie of the client. Jump to Step 2 for resource after login successfully. Otherwise, return login failure information directly.

5) The client calls user status query service with TokenID on Certification Center to judge whether the user is still in log state by visiting credential table. If it is, returns user name and user's authority to the data center and determines whether the user can access data resources by checking user's rights control. Otherwise, go to Step 4.

4.3 Web Services Based SOA

The basic demand of the distributed WSDSP is to implement data interoperate. As a new distributed interoperable model, web service can be used in different operation systems and integrate isomeric application systems easily. Therefore web service has become the mainstream of distributed interoperability technology and results in the occurrence of new service oriented architecture of software. SOA is an incompact software structure in which function models are encapsulated into web services and issues their interface onto internet. Developers only need call these interfaces and integrate them to build the software [5, 6].

For the advantages of SOA and the requirement of integrating and sharing distributed data, we adopt the SOA for WSDSP. This platform is comprised of one main center, one authentication center and several sub-centers. The architecture of this platform is shown in figure 3.

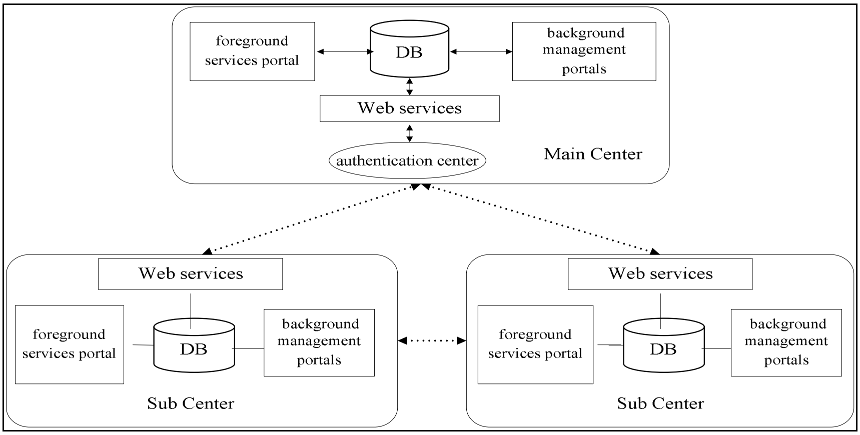

foreground services portal
DB
background management portals
Web services
authentication center
Main Center
Web services
foreground services portal
DB
background management portals
Sub Center
Web services
foreground services portal
DB
background management portals
Sub Center

Fig. 3. General architecture of WSDSP based on SOA

In "main center – sub center" structure mode, both of them require independent data service ability. That is, each of them has their own web portal, and can provide services such as user registration, information publish, metadata submission and collection, metadata censor, metadata inquiry, data browse and download, and so on.

The basic principle of SOA is to abstract a series of web services from the whole data sharing activities. These web services are respectively deployed in main center, authentication center and sub centers to implement the interoperability among them. The main center, authentication center and sub centers are both providers and consumers of web services.

As one node of WSDSP, besides providing services such as data submission, checkup and release, search and access, the main center is also the distributed business control center that harmonizes the interoperability among all centers. Sub centers can also implement the above functions independently, and communicate with each other through the main center. Authentication center offers services as user registration, login and authentication for all centers.

4.4 Distributed Metadata Synchronization

Metadata is the data about data [7]. It represents the who, what, when, where, why and how of the resource. By using metadata, people can publish, discover access and exchange data. Therefore metadata technology is adopted in many data sharing software [8, 9]. As a distributed architecture, in order to implement searching data one stop, WSDSP adopts a strategy of centralized managing metadata and decentralized storing data. In this strategy, all metadata should be collected in main center and the data can be stored in each sub center or corresponding holders.

In order to solve the problems of synchronizing metadata information between main center and sub center, we design the metadata synchronization service named MeteServ deployed in main center and each sub center. It includes existedMeta, addMeta, updaMeta, updaMetaState, deleMeta and getMeta interface. Through these interfaces we can synchronize the metadata of each sub center to main center. The detailed process for metadata submission and distributed synchronization is as follows:

1) The metadata will be stored in the local database of sub center after users' submission.

2) The administrator of this sub center censors the metadata, and if passed, call existedMeta interface on main center to judge whether the metadata has existed in main center. If main center doesn't have this metadata, it will call the addMeta interface to add a new metadata in main center; otherwise, call the updaMeta interface to update the same metadata in main center.

3) Then the administrator of main center censors the metadata synchronized from sub center and decides to revises, publish it or not. If it is revised, main center will automatically call the updaMeta interface of sub center to update the same metadata in sub center. Similarly if it is deleted in main center, it will be synchronizing deleted in sub center by calling deleMeta interface. If it is published, the updaMetaState interface of sub center will be called to change the metadata's state to public.

4) In order to ensure the data sharing service quality; the administrator of sub center and metadata creator can't delete public metadata directly, but they can revise it and request the administrator of main center to delete it. If the metadata is revised in sub center, it will be returned to Step 2 to be re-censored.

5 Conclusion

Adopting techniques and strategies mentioned above, we developed the WSDSP on J2EE environment and ORACLE10g; moreover, we used JBOSS as application server and AXIS-based web service as the communication mechanism between data centers. The platform can be divided into two modules, the foreground sharing system and the background administration system. The former was designed for data producers and users to provide the sharing services such as user register and login, metadata submission and collection, data directory navigation, metadata search and download and so on. While the latter one was designed for data management to provide data center management, users and authorities management, metadata management, statistical analysis and other information services. Up to now, we have finished the development of WSDSP and deployed one main center, one authentication center and 7 sub centers, in which more than 5876 MB data integrated from 5487 hydrometric station about 30 years can be shared to the public.

Acknowledgement. This work is supported by the Natural Science Foundation of China (No.51079040) and National Science and Technology Infrastructure of China (No.2005DKA32000).

References

1. Geng, Q.Z., Zhu, X.M.: Development of scientific data sharing of water resources and sharing service system construction in China. In: Proceedings of the 2010 IEEE International Conference on Software Engineering and Service Sciences (ICSESS), pp. 738–742 (2010)
2. Zhu, Y.Q., Feng, M., Song, J., Liu, R.D.: Research on earth system scientific data sharing platform based on SOA. Journal of Geo-Information Science 11(1), 1–9 (2009)
3. Zhang, Y.J., Xie, B.H., Guo, Y.Y.: The application of metadata technology in scientific data sharing platform. Journal of Taiyuan University of Technology 40(4), 341–344 (2009)
4. Zhu, Y.Q., Liu, R.D., Feng, M., et al.: Research on distributed earth system scientific data sharing platform. Computer Engineering and Applications 45(1), 245–248 (2009)
5. Wei, D., Chen, X.J., Fang, D.Y.: Research on Software Development Based on SOA. Micro-Electronics and Computer 22(6), 73–76 (2005)
6. Liang, A.H.: Thinking,Technology,System Integration and Application of SOA. Publishing House of Electronics and Industry, Beijing (2007)
7. Rust, G.: Metadata: the right approach,An integrated model for descriptive and rights metadata in e-commerce (2008), `http://www.dlib.org/dlib/july98/rust/07rust.html`
8. Goodchild, M.F., Fu, P.D., Rich, P.: Sharing Geographic Information: An assessment of the Geospatial One-Stop. Annals of the Association of American Geographers 97(2), 250–266 (2007)
9. Li, J.M., Xiong, A.Y.: Review of Meteorological Scientific Data Sharing System Research. Journal of Applied Meteorological Science 15(1), 1–9 (2004)
10. De Clercq, J.: Single Sign-On Architectures. In: Proceedings International Conference on Infrastructure Security, InfraSec 2002, pp. 40–58 (2002)

ExpertRec: A Collaborative Web Search Engine

Jingyu Sun[1,2], Junjie Chen[1,⋆], Xueli Yu[1], and Ning Zhong[2]

[1] College of Computer Science and Technology, Taiyuan University of Technology,
Taiyuan, Shanxi, China, 030024
[2] International WIC Institute, Beijing University of Technology,
Beijing, China, 100022
{sunjingyu,chenjj,yuxueli}@tyut.edu.cn, zhong@maebashi-it.ac.jp

Abstract. ExpertRec is a collaborative Web search engine, which is differ from current main search engine and allows users share search histories through a Web browser toolbar or a proxy browser. In addition, it can be taken as a novel social Web search engine and utilize expert's search histories for building recommendations. In this paper, we give an anatomy of ExpertRec and specially introduce its architecture and core techniques. It includes two basic components: a client agent and a backend server. The former is implemented as a Mozilla Firefox toolbar (a Firefox extension), which can integrate with mainstream search engines like Google, Yahoo!, et al., to meet users' teamwork needs. And it allows users to generate high-quality tags, votes, comments over current Web including search histories, personal archival content in local host typically beyond the reach of existing Web 2.0 social tagging system. The latter is a CBR (case-based reasoning)-based recommendation engine and implemented according to some core techniques, such recommendation rules, a scalable method to identify search expertise based on a hierarchical user profile in order to improve users' search quality, and so on. Finally, we give an evaluation and make conclusions.

1 Introduction

Web search has become one of the prominent information behaviors. However, current mainstream search engines and Web browsers are designed for solo use. But collaborative Web search(CWS) behaviors often occur in some search tasks, i.e. travel planning, literature search, technical information and so on. Recently, some demo CWS engines, i.e. Heystaks [7], SearchTogether [6], S^3 [5], CoSearch [1] and so on, are developed to study fundamental models and core techniques. Now CWS is becoming a staple way to improve search quality by users' collaboration[9].

In this paper, we introduce a novel collaborative/social Web search engine, ExpertRec, designed by us in details. Firstly, it can integrate with mainstream search engines like Google, Yahoo!, Bing, Baidu, et al. through a Web browser toolbar or a proxy browser; secondly, it allows every user to tag, vote, and share

[⋆] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 385–392, 2011.

his/her search histories and related Web pages; thirdly, it can recommend search expertise and re-rank search returns utilizing experts' search histories and so on. Current version depends on a Mozilla Firefox toolbar to catch search histories and display recommendations when a user searches in a Firefox Web browser. A CBR-based recommendation engine is designed to meet recommendation needs and some technique are adopted.

The rest of the paper is organized as follows. The next section introduces the architecture and interface of ExpertRec. Then, main techniques and core algorithms of ExpertRec are discussed in details. Finally, a evaluation is presented, and main related work is reviewed.

2 System Overview

2.1 Architecture

As shown in Figure 1, ExpertRec includes two basic components: a client agent and a back-end server. Any user has the alternative of a proxy browser or a client side browser toolbar as a client agent. In current version, the toolbar is implemented as a Firefox extension, which can integrate with Google search engine and so on, and a chrome extension version is being developed. The proxy browser is specially designed for two mobile platforms including Android and Apple and will be developed in future version. The back-end server includes a content sever, a CBR-based recommendation engine and a Web portal.

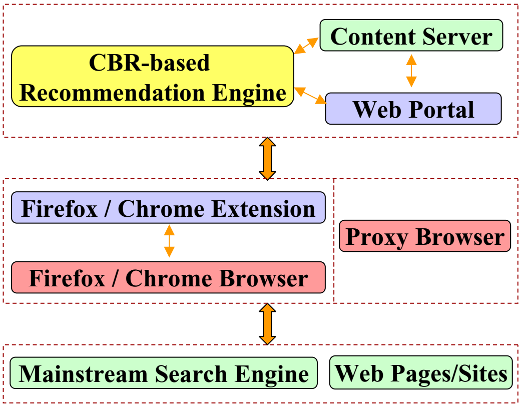


Fig. 1. System Architecture of ExpertRec

2.2 ExpertRec Toolbar

To support collaborative Web search, the ExpertRec toolbar is similar to Heystaks toolbar and depends five components shown in Figure 2: an ExpBase list, a group of buttons for tagging, sharing and voting Web pages, a drop-down menu about

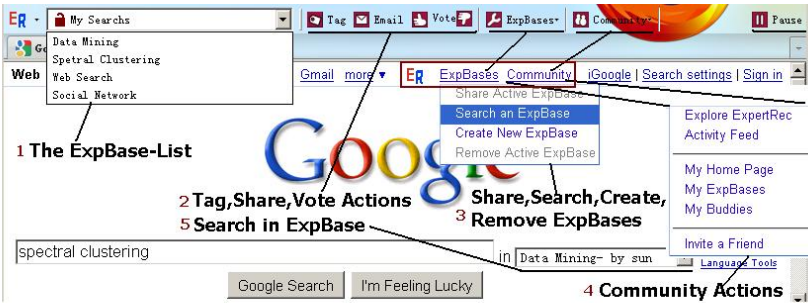


Fig. 2. Main functions of the toolbar when some logins

ExpBase, a drop-down menu about community and a start/pause button. Usually, every registered user can create several ExpBases for himself/herself with two types: private and public. The private ExpBase is only used to record his/her clicks in the Google result-list, tagged Web pages and so on in order to remind him in the future, but search histories in the public ExpBase can be shared with other users. Additionally, every user can invite his/her friends to join ExpertRec and share his/her search histories with them.

In addition, search results from default search engine are re-ranked through utilizing search expertise about one topic identified by our proposed method. And anyone can search experts' search histories when he/she chooses a ExpBase and types a keyword into input box of "recommendations search interface".

2.3 Back-end Server

The content sever is used to store search histories uploaded by the ExpertRec toolbar, search cases extracted from search histories, recommendations built according to our proposed method and so on.

The recommendation engine summarizes search cases and builds recommendations according to three recommendation rules in the following.

- First rule is that a recent search case associated with the query and appeared in his/her profile will be prompted in first place of promotion-list and used to remind him/her in order to avoid to browse repeatedly.
- Second one is that a search case with a largest *expScore* value is recommended in second place of promotion-list and is taken as his/her a possible interesting new Web page.
- Third rule is that a re-ranking of search results list is provided and returns visited before are marked for reminding.

The Web portal allows any registered user to login and logout, to manage their interests, to maintain search histories, to add/delete friends and so on. In addition, it provides an interface to search experts' search histories in a chosen ExpBase. For administrators, it provides some analyzing functions for tags, votes and so on through some effective data mining methods, such as spectral clustering [10], peculiarity oriented mining [11].

3 Core Techniques

3.1 CWS Environment and Search Results Extraction

With the help of Web browser plug-in or extension technology, we can design a CWS environment depending on a special toolbar, such as ExpertRec toolbar. In addition, we can capture search behavior (click action) and extract the title, url and others of the Web page or a search result in return-list by a search engine through programming Javascript[1] code. In addition, remote interactive techniques, i.e. remote conference systems, and instant messenger, such as QQ, MSN can be used to enhance CWS environment. And a special ExpertRec remote interactive component will be developed in the future.

3.2 Search Histories Representation

In ExpertRec, we designed a super case bases, named $ExpDB$, which includes m ExpBases. An ExpBase E is a case base related to one search task or topic, and is denoted by a set $E = \{expbaseid, creatorid, name, tags, type, description\}$. Usually, an ExpBase includes k search cases: $\{c_1, c_2, ..., c_k\}$ and a search case c is a summary of a Web page visited (search histories) including its title, queries, tags, votes, URLwords, snippet and selected-frequency. And it is denoted by $c = \{title, queries, tags, votes, URLwords, snippet, selected-frequency\}$. Specially, $snippet$ denotes its abstract or description; $URLwords$ denotes the keywords extracted from its URL; $select - frequency$ denotes the total number of times it is selected.

In nature, ExpertRec is a CBR system and provides a two-level case bases for managing search cases. Every public ExpBase is usually created by a trusted expert, and the system allows other users to join it and share its search cases.

3.3 Search Case Retrieval

According to Jaccard similarity [4], the similarity between a query q and a search case c is computed by

$$Sim(q, c) = |q \cap c|/|q \cup c|. \tag{1}$$

Where, q and c are taken as two list of terms. Specially, c only includes terms in its $title, queries, tags, URLwords, snippet$ through using some preprocessing steps like stop words removal and stemming. So all related cases to q in a given ExpBase E can be retrieved and ranked according to function 1.

3.4 Identifying Search Expertise

In order to identify search expertise, we adopt a topic filtering method proposed by us in [9]. The proposed method includes three key steps:

- to build a clustered hierarchical user profile in order to get the support of an topic (a term t), $Sup(t)$,

[1] http://en.wikipedia.org/wiki/JavaScript

- to define a filtering threshold $minFamiliar$ and take any term t in the user profile with $Sup(t)/|C| \geq minFamiliar$ as a user's familiar topic, and
- to take all familiar topics as search expertise for re-ranking Web pages.

The core of building the hierarchical user profile is the two heuristic rules: **similar terms rule** and **parent-child terms rule**. The former combines similar terms on the same interest and the later describes the parent-child relationship between terms. Based on them, the profile can be automatically built in a top-down fashion and represented by a tree structure, where each node is labelled a term t and associated with a set of supporting cases $S(t)$. In addition, the nodes corresponding to the familiar topics in the profile construct a expert profile U_{expert}, which is a connected subtree of the complete user profile stemming from the user profile root.

In ExpertRec, search expertise is transformed into a list of weighted terms: $< t, w_t >$, and the weight of each term in U_{expert} is estimated by applying the concept of IDF (Inverse Document Frequency). Given a term t, the weight of t, denoted by w_t, is calculated as:

$$w_t = log(|C|/Sup(t)). \tag{2}$$

3.5 expScore

According to information theory, the amount of information about a certain topic of the user is measured by its *self-information* [3]. For any term t,

$$I(t) = log(1/P(t)) = log(|C|/Sup(t)). \tag{3}$$

In order to estimate the value of a search case, we define an expert score, named *expScore*, for every search case. It is computed by $\sum_t I(t)$, where t denotes a term (a familiar topic) which appears in the queries and the search case at the same time. A search case with larger *expScore* is recommended preferentially.

3.6 Re-ranking

In ExpertRec, re-ranking list is built when a query is submitted to the recommendation server in five steps:

1. *Search expertise preparation:* The expert profile of every ExpBase is built and represented by a set of $< t, w_t >$ pairs in the recommendation engine server.
2. *Default ranked search results preparation:* The toolbar captures a query and the search results returned by a search engine, which are uploaded to the recommendation engine server for re-ranking. Each result comprises of a set of links related to the query, where each link is given a rank from the search engine, called $DefaultRank$.

3. *Computing EScore for ERank:* For each of the returned link l, a score called $EScore$ is calculated by the expert profile as follows:

$$EScore(l) = \sum_t w_t \times f_t, \tag{4}$$

where t is any term in the expert profile, and f_t is the frequency of the term t in the snippet of the link l. An $ERank$ is assigned to each link according to its $EScore$, and the link with the highest $EScore$ will be ranked first.
4. *Linear combination $DefaultRank$ and $ERank$:* Re-ranking results by combining ranks from both $DefaultRank$ and $ERank$. The final rank, $EERank$ (Expertise Enhancing Rank), is calculated as:

$$EERank = \alpha * ERank + (1 - \alpha) * DefaultRank, \tag{5}$$

where the parameter $\alpha \in [0, 1]$ indicates the weight assigned to the rank from the expert profile. If $\alpha = 0$, the expert profile is ignored, and the final rank is decided by the expert profile instead of the search engine when $\alpha = 1$.
5. *Recommending re-ranking list:* The toolbar downloads the final ranking of the search results and recommends them to the user.

The recommendation engine is implemented as a configurable platform in Java, and some other methods, such as expert finding [2] would be tested in future work.

4 Evaluation

In this paper, our main aim is to describe features and core techniques of ExpertRec. All experiments are conducted with following questions:

- As a CBR-based Web search recommender system, ExpertRec allows users to create multiple case bases (ExpBases) and share search histories with others. But do users actually take the time to create ExpBases and do they share them with others?
- As a search assistant, ExpertRec tries to improve Web search by facilitating collaboration among searchers. But do users benefit from this collaboration? Do they respond positively to ExpertRec recommendation? Do they benefit from their own search experiences or those of others or a mixture of the two?
- As a special expert system, ExpertRec allows common users to share search histories of experts. But How about the relationship between search quality and expertise identified by our proposed method?

In particular, we invite 20 participants who are chosen from different research groups in our labs to use and evaluate ExpertRec during the period October 2009-October 2010. They are with high levels of computer literacy and familiarity with Web search.

During the period of testing ExpertRec, about 30% participants create or join more than 10 ExpBases and about 30% participants only join ExpBases created by others. Furthermore, most of ExpBases store more than 40 search experiences. In total more 200 ExpBase were created and more than 8000 search experiences were produced. As a result, most of users are willing to utilize sharing features. In detail, We carry out an investigation about attitude for ExpertRec. 18 participants can accept the system to capture search histories, 16 like most of the function provided by ExpertRec Toolbar and have a good trial feeling. They are interested in some functions of toolbar, i.e. the vote (18 like), tagging (16 like), sharing by Email (15 like), and so on.

Through the creation and sharing of ExpBase, participants ought to find useful recommendations. According to access logs about clicked recommendations, we construct a user social network to show the relationship between recommendation producer and consumer who re-selects it. We analyze this network and find that re-selected recommendations come from about 80% users. We can conclude that ExpertRec can help users find their wants through recommending search experiences.

Furthermore, we conducted an experiment to explore the relationship between search quality and expertise identified by our proposed method and results shown in [9]. In our future work, we will explore another method to identify expertise through computing the out-degree for every user in the user social network and users with higher out-degree taken as expertise.

5 Related Work

To our knowledge, Heystaks[2] is only one similar system to ExpertRec introduced in [7]. In order to capture search experiences, ExpertRec implements a Firefox toolbar with similar features with Heystaks toolbar, but ExpertRec toolbar adds a few new features, i.e. search ExpBase, re-ranking recommendation. However, the main difference between ExpertRec and Heystaks is that ExpertRec adopts a novel method utilizing expertise. Furthermore, ExpertRec is been designed for mobile platform including Android and Apple.

6 Conclusions

Collaborative Web search is a promising way to improve search quality by users' working in cooperation. However, this approach requires a convenient way for users to work together. For this goal, we designed ExpertRec, a novel recommender system. It can utilize search expertise and integrates with mainstream search engine like Google via a browser toolbar. The toolbar allows users to tag, share and vote Web pages (including search histories and local Web pages). Search expertise is identified by analyzing a hierarchical user profile and recommended according three recommendation rules. Primary evaluation shows that ExpertRec provides some functions users like.

[2] http://www.heystaks.com

In our future work, we would extend it to support mobile search through developing a proxy browser in Android and Apple platform. In addition, we would extend CBR-based recommendation server through using a new case base maintenance approach for Web-scale CBR [8], and some new outlier detection algorithms [11] may be used to mine a possible interesting search case for users. Furthermore, evaluation will be thoroughly discussed.

Acknowledgment. The authors would like to acknowledge the following support for their research on ExpertRec: Youth Natural Science Foundation of Shanxi (No. 200821024); National Natural Science Foundation of China (No. 60873139); Open Fund of the State Key Laboratory of Virtual Reality Technology and Systems(No. SKVR-KF-09-04). And special thanks for Xianhua Li and PhD Jian Yang.

References

1. Amershi, S., Morris, M.: Cosearch: A system for co-located collaborative web search. In: CHI 2008, pp. 1647–1656 (2008)
2. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of SIGIR, pp. 43–55 (2006)
3. Cover, T.M., Thomas, J.A.: Elements of information theory, 1st edn. Wiley-InterScience, New York (1991)
4. Eduardo, D.: On clustering and evaluation of narrow domain short-text corpora. PhD thesis (2008)
5. Morris, M., Horvitz, E.: S3: Storable, shareable search. In: Baranauskas, C., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4662, pp. 120–123. Springer, Heidelberg (2007)
6. Morris, M., Horvitz, E.: Searchtogether: An interface for collaborative web search. In: IUIST 2007, pp. 3–12 (2007)
7. Smyth, B., Champin, P.: The experience web: A case-based reasoning perspective. In: Grand Challenges for Reasoning from Experiences, Workshop at IJCA 2009, pp. 566–573 (2009)
8. Sun, J., Yu, X., Wang, R., Zhong, N.: A model for personalized web-scale case base maintenance. In: Liu, J., Wu, J., Yao, Y., Nishida, T. (eds.) AMT 2009. LNCS, vol. 5820, pp. 442–453. Springer, Heidelberg (2009)
9. Sun, J.Y., Yu, X.L., Zhong, N.: Collaborative web search utilizing experts' experiences. In: Web Intelligence 2010, pp. 120–127 (2010)
10. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17, 395–416 (2007)
11. Yang, J., Zhong, N., Yao, Y.Y., Wang, J.: Local peculiarity factor and its application in outlier detection. In: KDD 2008, pp. 776–784 (2008)

A QoS Evaluation Method for Personalized Service Requests

Rutao Yang[1,2], Qi Chen[1,2], Lianyong Qi[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
{yrutao,lianyongqi}@gmail.com, adios737@126.com,
douwc@nju.edu.cn

Abstract. With the prevalence of Web service, QoS is playing a more and more important role in service evaluation, recommendation and selection. In most previous works, it is often assumed that the delivered QoS of a Web service is often determined by service provider, not service consumer. However, in the practical service execution environment, Web services usually work in an interactive mode with service consumer, so service consumer should also take responsibility for the delivered QoS of a Web service. Hence, it becomes a challenge to evaluate the QoS of Web services impartially. In view of this challenge, a QoS evaluation method for personalized service requests is proposed in this paper. Finally, the effectiveness of our method is validated, and an optimization method is proposed to improve the QoS evaluation efficiency.

Keywords: QoS evaluation, Service request, Collaborative Filtering, Clustering.

1 Introduction

Increments of personalized software requirements and the adoption of partial software outsourcing ideology bring new challenge for the traditional commercial off-the-shelf software solution. The Software as a Service (SaaS) model has become a possible solution by the widespread availability of fast Internet access, combined with widespread acceptance of service oriented architecture (SOA). However, when the number of software services that are delivered on-demand and priced on-use, has expanded dramatically, it becomes a challenge to select appropriate services for service requesters among many functionally equivalent services. Hence, evaluating the quality of service (QoS) [1][2] plays an important role in service recommendation, selection, and composition. Usually, the QoS published by service providers are distrustful, because QoS is variable during different service invocations and providers may give inauthentic high quality information in order to attract more potential service consumers. In this circumstance, the reputation of service providers or ratings of

[*] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 393–402, 2011.

services are mined based on historical execution QoS information. However, bad execution QoS is usually all ascribed to service provider's responsibility unfairly. A Web service is seen as an application accessible to other applications over the Web [3], which means services are provided in an attractive mode. Then personalized service requests will cause the variation of the service's delivered QoS. Most researchers pay more attention to the QoS evaluation from provider's perspective, and ignore the effect from consumer's perspective.

In view of the issue referred above, we put forward a QoS evaluation method for personalized service requests. Specifically, a service request model is proposed to specify personalized requirements, and then candidate services' quality are evaluated based on collaborative filtered historical execution information. The remainder of the paper is organized as follows: Section 2 discusses the motivation of this paper. Section 3 demonstrates our proposed personalized QoS evaluation method, followed by performance analysis and optimization of the method in Section 4. In Section 5, the related work is discussed. Section 6 concludes the paper.

2 Motivation

As explained before, QoS evaluation is an important step before service ranking, recommendation, selection and composition. Once the service provider-consumer relationship is potentially established, it usually refers to three kinds of QoS, which are published QoS (promised QoS), required QoS (expected QoS) and execution QoS (delivered QoS). They are time consecutive, and refer to three procedures of SOA (see Fig. 1). Along with a service is registered, QoS information is published to demonstrate the service's quality. Then consumers request a service with QoS constraints, and lastly consumers usually give feedback for the execution QoS after invoking a service. By the way, execution QoS information is monitored to record in historical log for helping service providers to improve service's quality. Published QoS is directly used for service selection or composition in some literature (say [4]). The trustworthiness of published QoS is ignored or published QoS is assumedly credible. However, in practice, the execution QoS cannot be known until the service is invoked and QoS evaluation is a necessary prelude.

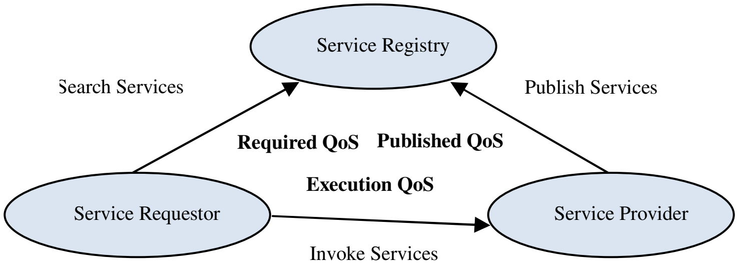


Fig. 1. Service Oriented Architecture (SOA)

The trustworthiness of published QoS is commonly measured by the service's reputation. As proposed in previous work, the reputation either depends on the ratings of user's experience [5], or is computed based on the actual measurement of the conformance of execution QoS to promised QoS [6]. However, the former solution suffers fairness problem of user ratings of services, especially in case malicious consumers may give false ratings and subvert services' reputation. And the later solution suffers from responsibility impartiality problem because of services are performed in an interactive mode. Consumers' network environment, personalized service request and payment will affect execution QoS. And empirical results show that different consumers likely have different experiences (e.g., Failure Probabilities) of invoking the same service (see Fig. 2). However, similar consumers' historical execution QoS can make a good contribution to the prediction of future execution QoS.

Based on the motivation discussed above, a personalized QoS evaluation method is proposed in this paper. Firstly, a service request model is defined for specifying personalized service requirement. Secondly, a collaborative filtering process of historical execution QoS is put forward. Then, the left records in previous step are used to compute the estimated execution QoS. Finally, the effectiveness of our method is validated, and an optimization method is proposed to improve the QoS evaluation efficiency.

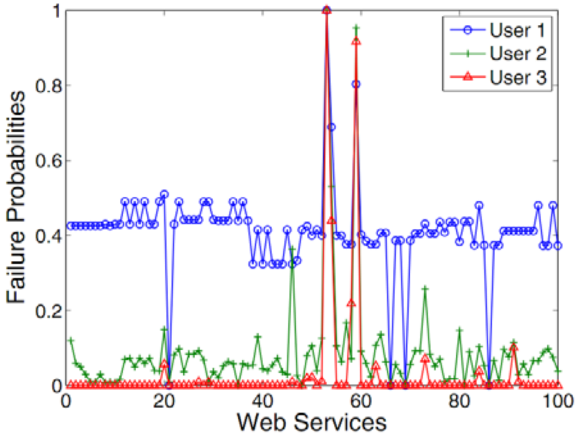


Fig. 2. Different users have different failure probabilities (from [7])

3 A QoS Evaluation Method for Personalized Service Requests

The goal of personalized QoS evaluation is to provide a fair and appropriate QoS evaluation in user-centric manner for service selection and recommendation.

3.1 Service Request Model

Usually, a service request contains functionality requirement and QoS constraints. In most work of QoS-aware service selection and composition, functionality requirement are used to discover candidate services that are functionally equivalent. However, the scale of input data or invoking times of a service will affect the execution QoS, especially execution duration. And the input data or invoking times are variable depending on consumers' requests. So, data input is extracted from traditional functionality requirement for specifying personalized service request, and the functionality still describes the ability of performing some computing task.

Fig. 2 shows that different users have different failure probabilities even when they invoke the same service. This phenomenon is caused by different context scenario, which characterizes the situation of a user, place or the interactions between users, applications and the environment [8]. As once a service request is delivered to a service provider, the main work responding the request is offered by the service. The interactions between service consumers and service providers are network environment, which affect the time in request delivering and result returning, especially while the service is about multimedia or large scale data processing. So, service requests are encouraged to contain client-side network environment, which can be measured by bandwidth, or other criterion.

As software services are priced on-use, payment should obviously be contained in a service request, and also payment is a main part of service level agreement (SLA). The reputation is a special criterion for evaluating a service, as it reflects the trustworthiness of a service provider (or published QoS), and then indirectly reflects a service's quality. So, in our service request model, the reputation is separated from other service quality, which is also used in literature [6].

Based on above discussion, a service request model is defined as follows:

Definition 1(Service Request Model): A service request is a tuple {*Functionality Description*, *Input Data*, *Network Environment*, *Price*, *Reputation*, *Quality Constraints*}, where *Quality Constraints* is also a tuple {*response time*, *availability*, *failure probability*}.

Specifically, *Network Environment* is not provided by service requester explicitly, but it can be evaluated based on the happening time and location of service request to mine network congestion. It's worth noting that *Quality Constraints* can be expanded by other QoS criteria. Take an online data mining software service for example, an instance of service request can be specified as {Mining task, 1G data, 100M bandwidth, 2$, above 9.0 (reputation's unbound is 10), {less than 10s, 95%,95%}}.

3.2 User-centric Collaborative Filtering

QoS evaluation is indeed a prediction of delivered QoS in future service invoking. Machine learning and reasoning under uncertainty have generated a variety of techniques that fall under the umbrella of predictive statistical models in Artificial Intelligence areas. Two main approaches have been adopted to perform prediction task:

content-based and collaborative [9]. In the application of QoS evaluation, the former approach requires that the service requester has ever invoked the service in history, which is usually unrealistic. So, collaborative approach, in which the QoS of a service is predicted from the behavior of other like-minded service invoking, is employed in this paper. As one of the most successful approaches to building recommender systems, collaborative filtering [10], in our QoS evaluation, uses the known QoS of a group of history users to make predictions of the unknown execution QoS for a new user.

As explained in Section 2, a collaborative filtering process is necessary for evaluating a service impartially in user-centric manner. Here the collaborative filtering is based on the factors that affect different users' experience QoS. As specified in Definition 1, functionality description is used for discovering candidate functionally equivalent services. Reputation and QoS constraints are used to service matchmaking. And the left elements in service request model, which can be demoted as a tuple named as *specialRequest*{*input*, *network*, *price*}, are the basis for collaborative filtering. The processing target of collaborative filtering is the historical records of service invoking. And the task of collaborative filtering is to identify the similar service requesters based on their similarity with a given *specialRequest*. Each service requestor' invoking information, demoted as *aRequest*{*input*, *network*, *price*}, can be extracted from its historical record. The similarity between *specialRequest* (*sR*) and *aRequest* (*aR*) is computed as follows：

$$similarity(sR,aR)=1-\frac{length(sR,aR)}{\sqrt{n}} \tag{1}$$

$$length(sR,aR)=\sqrt{\sum_{i=1}^{n}\left(\frac{|sR[i]-aR[i]|}{Max[i]-Min[i]}\right)^{2}} \tag{2}$$

Here *n*=3, as both *sR* and *aR* has three dimensions. *sR*[*i*] and *aR*[*i*] denotes the value of i*th* dimension of *sR* and *aR*. *Max*[*i*] and *Min*[*i*] denotes the maximum value and minimum value of i*th* dimension. The above formulas show that the *similarity* of *sR* and *aR* is based on their *length*, and the *length* is the normalized Euclidean distance. It's a general knowledge that the lower of their distance, the bigger of their *similarity*. The more *aR* is similar to *sR*, the more accurate (weight) its executive experience is supposed to be as a QoS prediction in next step. When a historical request' similarity with the new request is beyond of a threshold value, the corresponding invoking QoS information is filtered out, as it has less reference to the specialized QoS evaluation. The threshold value as a system parameter determines the number of left historical records. However, even the threshold value is set the maximum value, i.e., *sqrt*(*n*), the personalized QoS evaluation can also work on, because of the existence of weight, i.e., similarity, for each record.

3.3 Computing QoS Prediction

QoS evaluation can be categorized as the predictive statistical problem. And Section 3.2 has produced the filtered historical records set *S* and the weights *W* of these records for

QoS prediction. It's worth noting that a provider's QoS constraints are usually offered as an interval, and the candidate services for personalized QoS evaluation have satisfied the service matchmaking. As explained in Section 3.1, the criteria for personalized QoS evaluation contains *price*, *quality* {*response time*, *availability*, *failure probability*}, and *reputation*. All of these QoS criteria can be combined in a tuple *q* with size of 5, i.e., {*price*, *response time*, *availability*, *failure probability*, *reputation*}.

In mathematical statistics area, mean value and variance are two important criteria to represent estimation result. The corresponding formulas for computing QoS estimation in our method are denoted as follows:

$$E\left(q_i\right) = \frac{\sum_{i=1}^{m} w_i R_i}{\sum_{i=1}^{m} w_i} \tag{3}$$

$$V\left(q_i\right) = \sum_{i=1}^{m} w_i^2 \left(R_i - E\left(q_i\right)\right)^2 \tag{4}$$

where $E(q_i)$ and $V(q_i)$ are the mean value and variance of QoS criterion q_i, and m is the number of records, w_i is the weight of record R_i. It should be noted that the value of failure probability in each invoking record is 0 or 1.

As discussed in this section, a detailed algorithm is proposed as follows for concluding the personalized QoS evaluation method.

Algorithm. PersonalizedQosEvaluation(S)

Require: Service: S, Service Records: *Records*, Service Request: *ServR*, threshold value: *tv*

```
extract sR from ServR
extract Records from historical log about S
for all record $R \in Records$ do
    extract aR from R
    aR.weight = computeLength(sR,aR)      //(1),(2)
    if aR.weight <= tv then
        fR.add(R)
    end if
end for
for all criteria $q \in$ QoS do
    computeEq(fR);                        //(3)
    computeVq(fR);                        //(4)
end for
```

4 Performance Analysis and Optimization

In the following, the effectiveness and efficiency of our proposed method are both discussed, and an optimization method is devised to improve the efficiency of the method.

4.1 Effectiveness Analysis

The main purpose of our method is to evaluate each QoS criterion respectively for candidate services. Our method can be used before the overall quality evaluation of a service based on all these QoS criteria. Corresponding technique to compute overall quality has preference-oriented method [11], Simple Additive Weighting technique used in [5], and so on. Focusing on evaluation of single quality criterion, there are two methods for comparison, in which one is our method based on collaborative, and the other is the method used on [5] for computing quality criteria for elementary services. Take success rate for example, the value of the success rate is the ratio between the count of successful invocation and that of all history invocations without collaborative filtering in the later. As explained in Section 2 and experimental result showed in Fig.2, success rate is affected by both service provider and consumers. So our QoS evaluation method is more impartial for service provider and more accuracy to show service's quality to consumers. Even for some quality criteria that are not affected by service consumer, our method can still work on returning the same good value as former method as none historical QoS records is filtered.

The reputation is a special quality criterion that evaluates the trustworthiness of service providers. However, ensuring the veracity of reputation reports is a critical issue [6]. Focusing on above issue, one method is to compute the reputation by comparing the quality level that providers promise to the quality requirements. However, it neglects the consumers' subjective satisfactions for service usage, which is useful especially when consumer satisfaction for a single quality criterion is not linear proportional to the concordance between required QoS and delivered QoS. Our method can ensure the veracity of reputation reports by collaborative filtering malicious consumers' reputation based on similarity of service requests.

4.2 Efficiency Analysis and Optimization

Firstly, the candidate services set for personalized QoS evaluation can be reduced as small as possible by matchmaking with published QoS information. The matchmaking is based on the tenet that a service's actual quality cannot be better than the published quality. Non-skyline services are also not in the candidate services set. The count of candidate services is denoted as c. Then following considers one service for personalized QoS evaluation. Assuming the records number of execution log is r, the time complexity of collaborative filtering process is $O(r \times n)$, where n is the number of dimensions mentioned in formula(1) and (2). And the time complexity of predictive QoS computing is $O(m \times q)$, where m is size of left records and q is the size of QoS criterion. As $m \leq r$ and $n \leq q$, the overall time complexity of evaluating all candidate services is $O(\mathrm{c} \times r \times q)$.

Usually, q is limited, and is a constant for a set of functionally equivalent services. However, r is an increasing variable over time with the growing number of service invoking. And r can be an instable variable that causing bad performance of our method. A straightforward optimization method is cutting r by a time threshold, in which old records are not considered in QoS evaluation. We devise another optimization method based on user clustering, detailed as follows:

Step1: Cluster user requests of historical invoking using K-Means algorithm.

Step2: Find the representative user request for each cluster produced in ***Step1***.

Step3: Evaluate the QoS of the representative user request for each cluster using our proposed method.

***Step*4**: Response a service request, and compute out the nearest cluster of user request, followed returning the corresponding QoS evaluation result.

It is worth noting that ***Step1*▯ *Step3*** can be done offline periodically according the update of historical records, and ***Step4*** is done online for responding a new service request. Supposed the number of clusters is k, the time complexity of dynamic personalized QoS evaluating can be reduced to $O(k \times q)$, where k<<r.

The effectiveness and efficient of our method is verified by theoretical analysis. As a component of service selection system or service recommendation system, QoS evaluation cannot be experimented alone, our future work will be focus on applying our method into service selection for experimentation. In addition, our work has been in part ground on literature works in the area of SLA and QoS monitoring. SLA records a consumer's service requirement and QoS monitoring collects the historical execution QoS information.

5 Related Work

QoS is introduced into Web service in early literature [1], and [2], and the QoS criteria considered in these literatures contains availability, response time, throughput, security properties, accessibility, integrity, and regulatory and et al. QoS-aware service ranking, selection and composition get extensive research. Alrifai et al. [4] propose an efficient QoS-aware service composition method combining global optimization with local selection. Zeng et al. [5] described two service selection approaches, one based on local selection of services and the other based on global allocation of tasks to services using integer programming. Skoutas et al. [12] introduce service dominance scores based on multi-criteria dominance relationships for ranking and clustering Web services.

QoS computation or evaluation is a prelude of above work, and is concerned in literature [7], [11], [13] and so on. Rosario et al. [13] introduce soft probabilistic contract and use confidence interval and confidence to demonstrating service quality. Zheng et al. [7] evaluate real-world Web services by invoking these services in distributed manner and statistic QoS experience. However it is unrealistic for potential user to evaluate candidate services personally in this manner, and realistic method is estimate the QoS with historical invoking records. Liu et al. [11] extend QoS model with domain specific criteria and give a fair and open QoS computation method, in which all QoS information published by providers, from execution monitoring and requester's feedback are considered. However, it neglects the service consumer-side effect for execution QoS. Specially, trustworthiness or reputation of services is studied in literature [6], [14]. A method is developed for propagating reputation received by a composite service to its component services by Nepal el al [14]. Limam et al. [6] propose a reputation computation model based on automatic feedback computation for assessing software service quality and trustworthiness at selection time.

Similarly to our work, Thio et al. [15] introduce Client-Side Performance Estimation into Web service recommendation. The metrics related to client-side performance used in this paper are latency, transfer rate and throughput, which is similar to network environment in our proposed Service Request Model. Ivanovie et al. [16] address the issue of data-aware adaptation for service orchestrations. In our paper, a service request model is proposed for collaborative filtering and considers both network environment and input data, but also price, as higher price consumer pay, high quality provider provide. And a user-centric QoS evaluation method is devised for personalized service requests.

6 Conclusion

To handle the impartiality of QoS evaluation, a QoS evaluation method for personalized service requests is put forward in this paper. Specifically, a service request model is proposed to specify consumer-side affect for delivered QoS. Finally, the effectiveness of our method is validated, and an optimization method is proposed to improve the QoS evaluation efficiency. Our future work will focus on system implementing and experimental verification.

Acknowledgments. This paper is partly supported by the National Science Foundation of China under Grant No. 60721002, 61073032 and 60736015, and Jiangsu Provincial NSF Project under Grants BK2008017.

References

[1] Menasce, D.A.: QoS Issues in Web Services. IEEE Internet Computing 6(6), 72–75 (2002)

[2] Mani, A., Nagarajan, A.: Understanding Quality of Service for Web Service. IBM Developer Works (2002), https://www.ibm.com/developerworks/webservices/library/ws-quality.html

[3] Alonso, G., Casati, F., Kuno, H.A., Machiraju, V.: Web Services: Concepts, Architectures and Applications. Springer, Heidelberg (2004)

[4] Alrifai, M., Risse, T.: Combining Global Optimization with Local Selection for Efficient QoS-aware Service Composition. In: 18th International Conference on World Wide Web, pp. 881–890 (2009)

[5] 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)

[6] Limam, N., Boutaba, R.: Assessing Software Service Quality and Trustworthiness at Selection Time. IEEE Trans. on Software Engineering 36(4), 559–574 (2010)

[7] Zheng, Z., Zhang, Y., Lyu, M.R.: Distributed QoS Evaluation for Real-World Web Services. In: 8th International Conference on Web Service, pp. 83–90 (2010)

[8] Brezillon, P.: Focusing on context in human-centered computing. IEEE Intelligent Systems 18(3), 62–66 (2003)

[9] Zukerman, I., Albrecht, D.W.: Predictive Statistical Models for User Modeling. User Model. User-Adapt. Interact. 11(1-2), 5–18 (2001)

[10] Su, X., Khoshgoftaar, T.M.: A Survey of Collaborative Filtering Techniques. Adv. Artificial Intellegence (2009)

[11] Liu, Y., Ngu, A.H.H., Zeng, L.: QoS Computation and Policing in Dynamic Web Service Selection. In: 13th Internation Conference on World Wide Web, pp. 66–73 (2004)

[12] Skoutas, D., Sacharidis, D., Simitsis, A., Sellis, T.: Ranking and Clustering Web Services Using Multicriteria Dominance Relationships. IEEE Trans. on Service Computing 3(3), 163–177 (2010)

[13] Rosario, S., Benveniste, A., Haar, S., Jard, C.: Probabilistic QoS and Soft Contracts for Transaction-Based Web Services Orchestrations. IEEE Trans. on Service Computing 1(4), 187–200 (2008)

[14] Nepal, S., Malik, Z., Bouguttaya, A.: Reputation Propagation in Composite Services. In: 7th International Conference on Web Service, pp. 295–302 (2009)

[15] Thio, N., Karunasekera, S.: Web Service Recommendation Based on Client-Side Performance Estimation. In: Proceedings of the 2007 Australian Software Engineering Conference, pp. 81–89 (2007)

[16] Ivanovie, D., Carro, M., Hermenegildo, M.: Towards Data-Aware QoS-Driven Adaptation for Service Orchestrations. In: 8th International Conference on Web Service, pp. 107–114 (2010)

Virtual Personalized Learning Environment (VPLE) on the Cloud

Po-Huei Liang[1,2] and Jiann-Min Yang[2]

[1] Innovative DigiTech-Enabled Applications & Services Institute,
Institute for Information Industry, Taipei Taiwan
lph@iii.org.tw
[2] Department of Management Information System, National ChengChi University,
Taipei Taiwan
jmyang@mis.nccu.edu.tw

Abstract. With the virtualization technology maturity and the growing up fast of the cloud computing services, the application service providers have changed the way to server customers. Meanwhile, the new service model let users access the resource with the browser of their thin devices. By the way, the application service providers do not need to buy many machines for the uncertain backup or expanded requirement. Because the Cloud Service can provision the computing resources for customers dynamically, the users also pay as their use. The innovation of information science and technology drives e-learning system to produce a new type of service and the personalization requirements of the users are fast growing up. This paper presents a solution for building a virtual and personalized learning environment which combines the technology of Cloud Infrastructure as a Service (IaaS) and Cloud Software as a Service (SaaS) to create a service oriented model for the application service providers and the learners.

The proposed environment "Virtual Personalized Learning Environment" is intended for subscribing and excising of the selected learning resources as well as creating a personalized virtual classroom. This VPLE system allows the learning content providers to registry their applications in the server and the learners integrate other internet learning resources to their learning application pools.

Keywords: Cloud computing, e-learning, virtual personalized learning environment.

1 Introduction

Cloud computing is an extension of this paradigm where the capabilities of applications are exposed as services. These services enable the development of scalable web application in which dynamically scalable and often virtualized resources are provided as a service over the Internet [10, 16]. It is true that many international application providers such as Amazon, Google, IBM, Microsoft, and Sun Microsystems have begun to establish new type of data centers for hosting Cloud

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 403–411, 2011.

computing applications to provide redundancy and great reliability for meeting their Service Level Agreement. Since user requirements for cloud services are varied, service providers have to ensure that they can be flexible in their service delivery. Cloud computing makes it possible for almost anyone to deploy tools that can scale on demand to serve as many users as required [10].

The purpose of this paper is to present a cloud based solution for building a personalized learning environment for education. The rest of this paper organized as follows: Section 2 gives an overview of several kinds of learning environments and the critical factors proved by the TAM. Section 3 describes the characteristics and the advantages of the Cloud Computing. Section 4 presents the proposed Virtual Personal Learning Environment and the prototype system. Finally, Section 5 ends this paper with conclusion and future work.

2 E-Learning Environment

E-learning is defined as an Internet-enabled learning [18]. Components of e-learning can include content of multiple formats, management of the learning experience, and an online community of learners, content developers and experts. The study summarized the main advantages, which include flexibility, convenience, easy accessibility, consistency and its repeatability. Among the learning technologies, the Web-based learning offers the following benefits over conventional classroom-based learning including: (1) it can be used at any time and place; (2) the learning material is easy to update; (3) it fosters the interaction between the learner and the teacher in several ways; (4)it can incorporate multiple media such as text, audio, graphics, video and animation; (5) it enables learners to form learning communities; (6) facilitators can easily check learners progress, and (7) it allows for a learner-centered approach that can address the many differences between learners [15].

There are three main strategies in the fields of e-learning: Virtual Learning Environment (VLE), Personal Learning Environment (PLE), and Network Learning Environment (NLE).

2.1 Virtual Learning Environment (VLE)

A Virtual Learning Environment (VLE) is an electronic platform that can be used to provide and track e-learning courses and enhance instruction with online components [7, 18]. The VLE can automate the administration of learning by facilitating and then recording learner activity and it has evolved quite differently for formal education and corporate training to meet different needs. The contents are developed by teachers, which are mainly experts of a special domain. The VLE provides an easy to use system for flexibly delivering learning materials, activities, and support to students across an institution. For the administrator, a VLE provides a set of tools which allows course contents and students to be managed efficiently and provides a single point of integration with student record systems [7].

Nowadays, the trends of e-learning are to put emphasis on learner-personalized learning. This kind of learning places learner at its heart. Learners are expected to actively engage in the process to construct their own learning contents. Thus they would pay more intensions for their learning [9, 13].

2.2 Personal Learning Environment (PLE)

Applying Web 2.0 technologies to e-learning can enhance the interactive communication and the collaboration, or can help to discover and obtain the resources, or are able to exchange and share the resources with others. Thus, Web 2.0 provides a learning environment have the potential to fundamentally change the nature of learning and teaching, through the creation of learner controlled learning web. This learning environment is named a Personal Learning Environment (PLE) [13].

There are several reasons for adopting PLE as the platform for e-learning. The most important reason is that the PLE can help learners control and manage their learning contents. A PLE also permits learners to join into other learning groups and provides a suitable environment to practice social skills. Furthermore, the PLE can provide support for lifelong learning that is mainly informal and occurs over the life of the learner [13].

2.3 Network Learning Environment (NLE)

A Networked Learning Environment (NLE) is about the learning that takes place in a vibrant community of people and resources [19]. The Internet has removed the limits of time and proximity that once restricted this community. Similar to the Internet, a Networked Learning Environment is really a network of networks and the power of the NLE today is that it creates more possibilities for students and faculty, far beyond the limitation of books, bricks and mortar [20].

Critical Factors of e-learning systems.

Many researches proved the technology acceptance model (TAM) to be a good theoretical tool to understand users' acceptance of e-learning [1-8]. E-learning self-efficacy was one of the important construct, followed by subjective norm in explicating the causal process in the model [6]. The success of Web-based learning depends on learner loyalty, and the results indicated that performance expectancy, effort expectancy, computer self-efficacy, attainment value, utility value, and intrinsic value were significant predictors, should be examined at the same time [2, 5]. In other studies, the research results indicate that perceived usefulness has a direct effect on VLE use. Perceived ease of use and subjective norm have only indirect effects via perceived usefulness. Both personal innovativeness and computer anxiety have direct effects on perceived ease of use only [4, 7].

The proposed VPLE is developed on these critical factors:

- Ease use of learning systems
- Stability and Performance of learning systems
- Perceived ease of use
- Perceived usefulness

3 Overview of Cloud Computing

Cloud computing provides the new kinds of on-demand information technology services and products [26]. According the related studies of cloud computing [21-25], several most characteristics of the cloud computing are summarized as the following descriptions: (1) On demand service, (2) Ubiquitous network access, (3) Location-independent resource pooling, (4) Rapid elasticity, and (5) Pay per use.

In order to meet the characteristics of the cloud service, the fundamental action is to make resources virtualized [26]. Thus we can find that data centers can dynamically "provision" on demand to meet a specific service-level agreement [27, 28].

The types of on-demand Cloud service in Table 1 are respectively referred to as Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) [21, 27, 28, 29].

The other critical technologies for cloud resource management, conducted in our proposed platform, are described as follows [10, 16]:

1. To provide the dynamic scale-in and scale-out of applications by the (de-) provisioning of resources (for example, via virtualization).
2. To provide the monitoring of resource utilization to support dynamic load-balancing and reallocations of applications and resources.

Table 1. Three types of basic cloud services

Type	Description
Infrastructure as a Service (IaaS)	• Deploy the virtualized hardware resources as a service. • Users do not need to buy the server, network equipment, and storage equipment, only need through Internet lease can put up one's own application system.
Platform as a Service (PaaS)	• Offer application platform for internet programming interface, operation platform, and etc. • Users can structure and dispose their own application by the platform.
Software as a Service (SaaS)	• Customers use the software service on Internet through standard Web browser. • Provide multi-tenancy support. • Customers needn't buy the software, only need to rent the software as required.

4 Virtualized Personalized Learning Environment (VPLE)

Many e-learning and educational institutions are beginning to take advantage of existing applications hosted on a cloud that enable users to perform tasks that have usually required site licensing, installation, and maintenance of individual software packages [11,12,14,17]. Many applications such as word processing, spreadsheets, presentations, databases and more can all be accessed through a web browser, while the objects are housed in the cloud. Furthermore, it is very easy to share content created with these tools, both in terms of collaborating on its creation and distributing the completed work. In the current online Application Markets, such as Apple App

Store for iPhone, Mac App Store, and Android App market, the applications are ease to be put into the application server by the owners and the users can subscribe and get the application software on demand through the network. The users can create their personalized software environment and the application creators can get the incomes from subscribed users' payment. This successful service model is proved by markets and there are more service providers willing to conduct this service model for their business.

Virtual Personalized Learning Environment (VPLE), proposed in this paper, is a web-based platform that is built on the Cloud, and conducts a service oriented mechanism of the learning environment to serve the learners and the content providers. Based on the browser-based applications can be accessible with a variety of computer and even mobile platforms, making these tools available anywhere the Internet can be accessed. The purpose of this research was to develop a solution which integrates multiple disparate application systems of schools into one integrated online system that may be accessed via a single sign-on process. The intent of this system is to utilize new web-based technologies to enhance teacher performance and student learning.

4.1 Architecture of VPLE

Figure 1 shows the architecture of the proposed Virtual Personalized Learning Environment (VPLE). The system combines the virtual resource management functions of the Cloud IaaS and the personalized learning management system in the Cloud SaaS.

- The functions used in the Cloud IaaS: (1)Storage management for the learning system and the users, (2)Load Balance for all learning systems, (3)Scaling management for virtual machines, and (4)Backup and Restore for the learning applications.
- The functions used in the Cloud SaaS: (1)Application Registry management for the commercial provides to register their applications, (2)Application Server for managing and deploying the subscribed learning contents to the users, (3) Account manage system for the authorized users, (4)Virtual Desktop Deployment for providing the personalized desktop including the subscribed learning contents, (5)Session Management for ensuring the Virtual Desktop used by the authorized user, and (6)Personalized management for managing the subscription of the favorite learning contents.

4.2 Service Model of VPLE

In order to indicate the difference of the VPLE from other PLE and VLE, the service model and the critical process flows are illustrated in Fig. 2. The following three process flows explicate the core ideas of the application store like and the personalization in the VPLE. The first process flow is that the commercial application manager or the teacher would like to register a new application into the application server. The second process shows the authorized learner how to get the virtual desktop. At last, the process of the Personalization emphasis on users can select the additional learning contents for their favorites.

Based on this service oriented model, the learning content providers can dynamically register new applications at any time and users can personalize the contents of their learning environments. Further, all systems in the VPLE are virtual machines. If the performance of the system became slower than the standard of the SLA, the load balance system of the hypervisor would be enabled to share the workloads with other virtual machines of the same application.

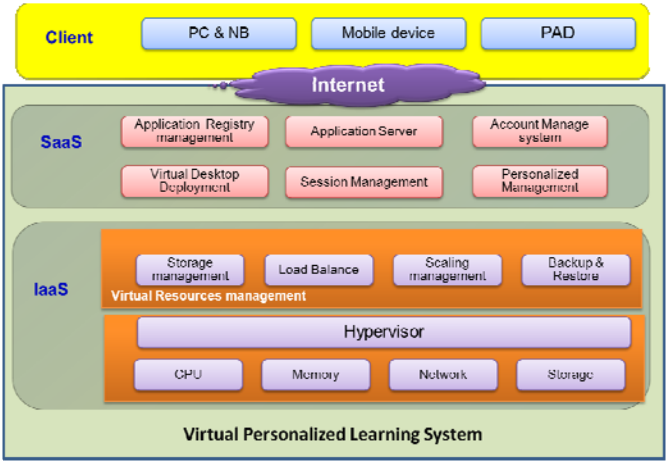


Fig. 1. The Architecture of the Virtual Personalized Learning Environment

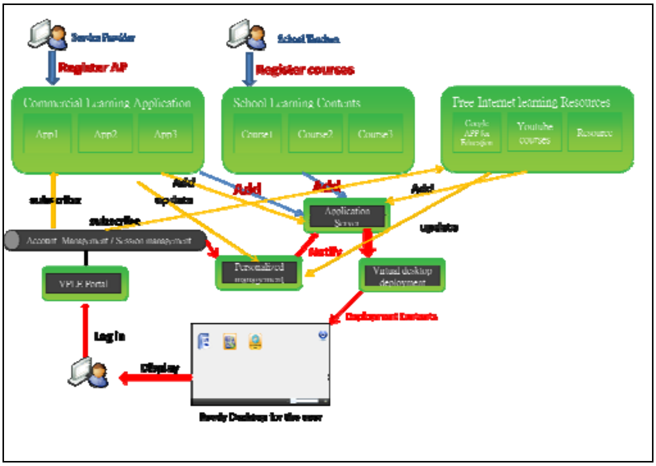


Fig. 2. The Service Model and the Process Flow of the VPLE

4.3 Implementation and Evaluation of Performance

The prototype system was established and ever demonstrated for the teachers of the primary school and the junior high school. In our environment, there were four virtual machines created for the VPLE.

The next demonstration is to explore the performance of the portal virtual machine. The core hardware configuration of the portal virtual machine is 4 vCPUs and 2048 mega RAM. The stress tool used is Pylot and it is an open source web performance tool. The stress testing was conducted with several agent parameters and the testing results were gathered in Table 2.

Each testing agent would run with the same configuration which includes the duration of sixty seconds, the ram up time of zero second, and the interval time of zero millisecond. As the analysis of the test results, there are several findings for the future improvement: (1) the CPU utilization was increased with the numbers of the testing Agents, (2) the upper bound of the successful requests would be smaller than 4,900, and (3) the average response time can be provided to the VPLE service provider for the Service Level Agreement.

Table 2. The results of the web performance testing

Testing Agent #	**Testing results**			
	CPU ult	Total Requests	Avg Resp time(sec)	Throughput (req/sec)
1	24%	1250	0.048	20.492
10	75%	4750	0.124	78.230
50	67%	4684	0.630	75.548
100	59%	4580	1.291	75.082
120	69%	4710	1508	76213
130	65%	4649	1656	76.213
135	78%	4874	1.646	78.613
140	69%	3392	1.924	69.079

5 Conclusion

Ease use and system performance of e-learning system are of vital importance to the learners [3, 4]. From IT perspective, cloud computing is a cost down and efficient IT management technology for the application service providers [1-5]. Certain cloud enabled investments that will have implications for version migration and new application development.

In this paper, we investigate how the new service models brought about by cloud computing enhance the improvement on the efficiency within the VPLE. From elastic infrastructure aspect, the capabilities and implementation mechanism of cloud SaaS based e-learning solution is presented. Finally, the performance and capacities of the VPLE portal site is tested by Pylot and the testing results are analyzed at the same time.

Although the VPLE has been proved of concept, there are many aspects need to be improved. Our future works will focus on how to load balance the workloads from the large users over the internet and how to make users feel ease use.

Acknowledgments. This study is conducted under the "The Core Technology Research and Development for Cloud Application Project" of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

1. Saade, R.G.: Web-based educational information system for enhanced learning. EISEL: Student assessment. Journal of Information Technology Education 2, 267–277 (2003)
2. Chiu, C.-M., Wang, T.G.: Understanding Web-based learning continuance intention: The role of subjective task value. Information and Management 45(3), 194–201 (2008)
3. Stoel, L., Lee, K.H.: Modeling the effect of experience on student acceptance of Web-based courseware. Internet Research 13(5), 364–374 (2003)
4. Davis, Fred, D.: User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. International Journal of Man-Machine Studies 38(3), 475–487 (1993)
5. Liaw, S.S., Chen, G.D., Huang, H.-M.: Users' attitudes toward Web-based collaborative learning systems for knowledge management. Computers & Education 50(3), 950–961 (2008)
6. Park, S.Y.: An Analysis of the Technology Acceptance Model in Understanding University Students Behavioral Intention to Use e-Learning. Educational Technology & Society 12(3), 150–162 (2009)
7. van Raaij, E.M., Schepers, J.J.L.: The acceptance and use of a virtual learning environment in China. Computers & Education 50(3), 838–852 (2008)
8. Ong, C.S., Lai, J.Y., Wang, Y.S.: Factors affecting engineers' acceptance of asynchronous e-learning systems in high-tech companies. Information & Management 41(6), 795–804 (2004)
9. Popescu, E.: Adaptation provisioning with respect to learning styles in a Web-based educational system: an experimental study. Journal of Computer Assisted Learning 26, 243–257 (2010)
10. Weinhardt, C., Anandasivam, A., Blau, B., Stosser, J.: Business Models in the Service World. IT Professional, 28–33 (2009)
11. Basal, A.M., Steenkamp, A.L.: A Saas-Based Approach in an E-Learning System. Iranian J. Information Sci. Management, Special Issue, 27–40 (2010)
12. Al-Zoube, M.: E-Learning on the Cloud. International Arab Journal of e-Technology 1(2) (2009)
13. Harmelen, M.: Design trajectories: four experiments in PLE implementation. Interactive Learning Environments 16(1), 35–46 (2008)
14. Xiao, L., Wang, Z.: Cloud Computing: A New Business Paradigm for E-learning. In: International Conference on Measuring Technology and Mechatronics Automation, pp. 716–719 (2011)
15. Jolliffe, A., Ritter, J., Stevens, D.: The online learning handbook: Developing and using Webbased learning. Kogan Page, London (2001)
16. Hutchinson, C., Ward, J., Castilon, K.: Navigating the Next-Generation Application Architecture. IT Professional 11(2), 18–22 (2009)
17. Rajam, S., Cortez, R., Vazhenin, A., Bhalla, S.: E-Learning Computational Cloud (eLC2): Web Services Platform to Enhance Task Collaboration. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp. 350–355 (2010)

18. Gunasekaran, McNeil, R.D., Shaul, D.: E-learning: research and applications. Industrial and Commercial Training 34(2), 44–53 (2002)
19. Wei, H., Wang, F.: Application of Cloud Computing in the Network Learning Environment. In: International Symposium on Computational Intelligence and Design, pp. 205–208 (2010)
20. Pittinsky, M.: The Networked Learning Environment. Stepping Beyond Courses to a More Expansive Online Learning Experience (2004)
21. Linthicum, D.S.: Cloud Computing and SOA Convergence in Your Enterprise: A Step-by-Step Guide. Addison-Wesley Information Technology Series. Addison Wesley, Reading (2009)
22. Greer, M.B.: Software as a Service Inflection Point: Using Cloud Computing to Achieve Business Agility, iUniverse, Inc. (2009)
23. Menken, Blokdijk, G.: Cloud Computing Virtualization Specialist Complete Certification Kit - Study Guide Book and Online Course, Emereo Pty. Ltd. (2009)
24. Velte, T.: Cloud Computing: A Practical Approach. McGraw-Hill, USA (2009)
25. Rittinghouse: Cloud Computing: Implementation, Management, and Security, 1st edn. CRC Press, Boca Raton (2009)
26. Vouk, M.: Cloud Computing—Issues, Research, and Implementations. In: Proc. 30th Int'l Conf. Information Technology Interfaces, pp. 235–246. Univ. Computing Centre, Zagreb (2008)
27. Buyya, R., Ranjan, R., Calheiros, R.N.: Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities. In: Proc. of the 7th High Performance Computing and Simulation, Leipzig, Germany (2009)
28. Youseff, Butrico, M., Silva, D.D.: Towards a Unified Ontology of Cloud Computing. In: Grid Computing Environments Workshop, GCE 2008 (2008)
29. Motahari-Nezhad, H.R., Stephenson, B., Singhal, S.: Outsourcing Business to Cloud Computing Services: Opportunities and Challenges. Technical Report HPL-2009-23 (2009)
30. Pylot, `http://www.pylot.org`

MTrust-S: A Multi-model Based Prototype System of Trust Management for Web Services

Dunlu Peng, Shaojun Yi, Huan Huo, and Jing Lu

School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai, 200093, China
dunlu.peng@gmail.com, angle20161026@163.com,
jinglu76@gmail.com, huoh@usst.edu.cn

Abstract. As an important technique for internet-scale information integration, web service becomes popular very rapidly in recent years. More and more enterprises move their core business onto the Web in the form of web services. Malicious web services will affect the security and reliability of the requester's application which invokes the services. Therefore, identifying the trustworthy web services is now a critical issue to make requester's application secure and reliable. In this paper, we develop a prototype that naturally provides a solution for the evaluation and management of web service trust and reputation. The prototype integrates service requester's feedback collection, trust evaluation and trust management together. With collaborating of these three components, our prototype provides an effective way for selecting trustworthy services for the requesters. To model service trust more precisely, we present a mathematic presentation for different types of data describing service trust, i.e, discrete values, probabilistic values, brief values and fuzzy values. A series of models has been developed to evaluate the trust of services according to the types of trust values. Simulation results verify that using these models can greatly improve the success rate of invoking trustworthy services.

Keywords: trust management, web service, evaluation model, trust representation.

1 Introduction

With lots of advantages, such as loosely couple, well-defined contract, meaningful to requester, open, standard-based and some other characteristics, web service has been the de facto most popular technique for internet-scale information integration. In order to serve the clients better, more and more organizations move their core business onto the Web in the form of web services. Meanwhile, some malicious web services, which bring security concerns to requesters' applications or do some selfish behaviors, have also been published and intermixed with the trustable services on the Web. However, business applications, especially those needs to pay for the accessing information provided by web services, require the invoked services must be secure and reliable. In this work, we propose a framework for trust management of web services, and make the following contributions:

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 412–419, 2011.

1. *Representation of service trust metrics.* Generally, more than one metrics have been applied to describe the trust of service. These metrics may have different types, for example, some have discrete values and some have uncertain data. Our representation concerns both of them.
2. *Multiple evaluation models for computing the synthetically trust of service.* We propose a series of methods to compute the synthetically trust of service from different aspects. Users can select the appropriate methods to evaluate the trust of service according to their requirement.
3. *Algorithms for computing the trust relationship between services.* We employ group of algorithms through the combination of classical trust models with the service computing environment. The proposed models evaluate the trust relationship among services. To explain how to use these models, we conducted some experiments as the examples for evaluating performance of each algorithm.
4. *A Framework of the prototype of managing the trust of web services.* The framework named as *MTrust-S* describes the prototype of web service trust management. It is implemented in the Intranet of our laboratory.

In next Section, we give some introduction on the related work. In Section III, we describe some definitions and trust representation of service, trust inference models and evaluation functions. *MTrust-S*—the framework is proposed in Section VI and the experimental results are also shown in the same section. Finally, we draw the conclusions in Section V.

2 Related Work

With the popularity of the Internet, trust-based information system has received considerable attention in recent years. Much previous research has been done on trust managements in applications ranging from online auctions to peer-to-peer networks. We give an overview of these approaches as follows.

Policy-based models regard policies as a set of rules that specify the conditions to disclose own resources. In semantic web services, trust policies are formulated using security statements in many models, such as confidentiality, authorization, authentication [1].

Trusted Third Party-based models work as a repository of service description and policies, and also act as an external matchmaker that evaluates service trustworthiness using given algorithms. Some policy-based models rely on this kind of models [2].

Reputation-based models take use of rating coming from other agents or central engine by heuristic evaluations. These models reuse the concepts and approaches which are extracted from the Web-based social networks [3]. Trust management is a critical issue in service computing as well as in social networks.

Reputation management systems have also been developed for the clients having the requirement of selecting the most trustable web services. In [4], web service quality model in which trust is considered as one of the five quality attributes of services.

Extinguishing from the existing research on web service selection, our trust management prototype evaluates service trust from the multiple perspectives with different models, which can select the most trustable service for clients.

3 Service Trust Representation and Evaluating Models

In this section, we first present the definitions, representations and algorithms for computing trust degree of web services which are employed in our system.

3.1 Definitions for Service Trust

So far, there is no generally accepted definition of trust computing in the literatures. In our work, we descriptively define web service trust according to its characteristics as follows.

Definition 1. Web service trust depicts the degree of a service meeting a requester's requirements under the limitation of the requester's computing context. It is evaluated by the requester according to the invoking historical records of the service and the comment on the service from other requesters who have ever invoked it. From the source where the requester obtains the trust information, service trust could be classified into two kinds: *direct service trust* and *recommended service trust.*

Definition 2. Direct service trust refers to the service trust of service gained by the requester using its own historical recorders of invoking the service. If the trust of service is also obtain from the comments of other requesters, it is called as **recommended service trust**.

Definition 3. Service trust degree measures service trust quantitatively. The trust degree of service w to requester q is denoted as $T_{w\to q}$. More specifically, the direct trust degree and recommended trust degree are defined as $T_{w\xrightarrow{d}q}$ and $T_{w\xrightarrow{r}q}$ respectively.

Definition 4. Service reputation is the synthetical measurement of service trust or quality, and reflects the level of the service serving its requesters and how its requesters trust it.

3.2 Representation of Service Trust

Mathematical representation of service trust is the foundation of trust management of service. Several methods were proposed to represent the trust in social activities in the literatures [5, 6]. In our work, we employ some of them to describe service trust.

Rep I: Discrete value. Inspired by PET [7], we classify service into four categories according to their trust, which are formalized as a set T={C, L, N, U}. The services labeled as L, N, U are something untrustworthy and make the requester decrease the value of trust. We use function $T(x)$ as the map function which maps from T to a numerical value v:

$$T_w(x) = \begin{cases} v_1, & x = C, & v_1 > 0 \\ v_2, & x = L, & v_2 < 0 and \mid v_2 > v_1 \mid \\ v_3, & x = N, & v_3 < v_2 \\ v_4, & x = U, & v_4 < v_3 \end{cases} \tag{1}$$

Rep II: Probabilistic value. This approach uses probability to represent the service trust. Given a requester *q,* the service trust of service *s* to *q* is set to *p∈[0,1]*, where the service trust increases as *p* increases. When *p=0* means service *s* is untrustworthy to requester *q*, while when *p=1* means service *s* is complete trustworthy to requester *q*.

*Rep III: Belief trust value. B*elief trust value which will be used to represent service trust is based on a belief model which keeps the uncertainties hidden in the theory of probabilities [8]. We introduce '*opinion*' proposed in [9, 10] to express the service trust synthetically as

$$b+d+u=1 \quad b,d,u\in[0,1] \tag{2.1}$$

$$E(w)=b+au \tag{2.2}$$

in which *b, d* and *u* refer to trustworthy, untrustworthy and uncertain respectively. *opinion={a, d, u, a}* , *E*(*w*) is the service trust degree and *a* is the coefficient representing the uncertainties in proportion of service trust degree.

Rep IV: Fuzzy trust value As we know, service trust itself is fuzzy, here we adopt fuzzy approach to represent service trust. The membership degree $S(w_s)$ is regarded as the degree of service trust set *S* containing service w_s. Given fuzzy service trust sets $S_1, S_2,\ldots,S_n$, the degree of a service *ws* belongs to each fuzzy set can be described together as a vector $v_{ws}=\{v_1,v_2,\ldots, v_n\}$, where v_i is the membership degree of *ws* to fuzzy set S_i and satisfies $\sum_{i=1}^{n} v_i = 1$.

3.3 Service Trust Models

Corresponding to the trust inference models aforementioned, in the trust management system, we employ four trust models to manage service trust. These models are as follows:

Model I: eBay-like Trust Model (eBTM) The *eBTM* employed in our system is based on the trust model used in eBay. In the model, requesters evaluate the service after invoking. The evaluating information includes positive evaluation (+*1*), neutral evaluation (*0*), negative evaluation (*-1*) and a short comment which has no effect on the computation of trust. Suppose after an invoking, requester *q* evaluate the trust degree of service *w* as $T_{w\to q}$, then the total trust degree of service *w* is $\sum_{i=1}^{k} T_{w\to q_i}$ and *k* is the number of requesters.

Model II: EigenTrust-like Model (ETM) EigenTrust algorithm computes global trust degree from direct trust degree by exploiting the transitivity of trust [10]. Similar to EigenTrust, our *ETM* is also based on this idea.

Suppose the trust degree of service *w* to requester *q* is $T_{w\to q}$ after *q* invoking *w* for several times. In order to decrease effect of the malicious behavior, $T_{w\to q}$ needs to be normalized. We denote the normalized $T_{w\to q}$ as *C*. WE obtain

$$T^{(k+1)}=(1-\alpha)C^{\mathrm{T}}T^{(k)}+\alpha p \tag{3}$$

where $T^{(k)}$ is the vector of global trust value after *k* iterations, *p* is the global service trust degree and α is the coefficient.

Model III: Dirichlet-like Model (DLM) Suppose that there are *k* discrete trust level for services, then the potential of state space for Dirichlet distribution is *k* in the trust model.

Let M denote all the requesters evaluation on service *w*, then $T_{w,t}(i)= \sum_{Tw \to q,t \in M} T_{w \to q,t}(i)$ .After (t+1) time intervals, the cumulative evaluation is computed as

$$T_{w,(n+1)} = \lambda \cdot T_{w,n} + T_{w,(t+1)} \tag{4}$$

Where λ is the attenuation factor of requester's evaluation. The finial evaluation of trust degree for service w is represented as a vector $VT_{w,n}$.

Model IV: Fuzzy Trust Model (FTM) Our *FTM* is derived for Fuzzy-based Trust Evaluation proposed in [11]. Let *S* be the times of invoking, E_{val} be the invoking evaluation, t_c be the current time and t_e be the evaluating time, then *SWTV* will be obtained

$$SWTV = \frac{\sum_{s=1}^{S} \left[e^{-(t_c - t_e)/\lambda} ((E_{val} - E_{\min}) / (E_{\max} - E_{\min})) \times 5 \right]}{S} \tag{5}$$

4 Prototype System and Performance Simulation

In this section, we first give an overview of our prototype system –***MTrust-S***, and then we inspect the performance of the system using simulation.

4.1 Prototype System

In Internet-scale, many services are deployed and evaluating the degree of service trust and storing the trust information in an effective way can enhance greatly the performance of the service-based applications. We present a prototype named ***MTrust-S*** to manage the trust of service. ***MTrust-S*** is a multi-model based system to evaluate and mange service trust. The models and algorithms are present in above sections. ***MTrust-S*** includes three components: requesters' feedback collector, service trust evaluator and trust management core. These components are depicts in Figure 1.

Our prototype evaluates service trust is based on the feedback gathered from the service's previous requesters. The *Requesters' Feedback Collector* (RFC) is the mechanism we use to collect the requester's feedback which is observed from the requesters after they interact with the service. *Trust Evaluator* (TE) whose role is to employ the service trust models and algorithms to compute the trust of services according to the information collected by the requesters' feedback collectors. *Trust Manager* (TM) is the center of the prototype and is responsible to manage services' trust, including storing trust, indexing trust querying trust and etc. We present the detailed description of each component as follows.

Requesters' Feedback Collector (*RFC*). Identifying the requesters who invoked the evaluated service and collecting the comments of these requesters on the service trust is the responsibility of RFC. Let *Q* be a service requester having little knowledge on the trust of the evaluated web service *S*. In order to obtain the correct evaluation of *S, Q* needs to get more comments on *S* from other requesters. Some ways can be exploited to collect the comments on *S*. In our prototype, *Q* broadcasts a question as the survey to all the requesters in the network. The question is 'Have you ever invoked service *S*? If so, what is your comments on it?'. If the requester receives the broadcast and has called S, it will send its comments on service *S* to *Q*. In this way, *RFC* collects the feedback of trust for all the services that need to be evaluated.

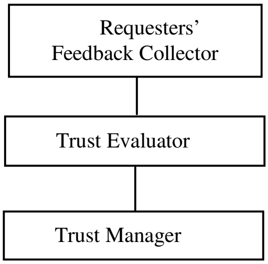


Fig. 1. Components of the Prototype

Trust Evaluator (*TE*). *TE* is responsible for computing the trust with the model presented in Section III according to the direct and indirect information gathered by *RFC*. *TE* is the kernel of our prototype and determinates the effectiveness of the whole system. An important issue *TE* should consider is how to process the feedback and how to combine the information obtained directly from the service and the information gathered from other requesters to form the final evaluation trust degree. The final evaluation trust degree will provide basis for decision-making. Generally, requester likes to choose the web service with maximum final evaluation trust degree.

Trust Manager (*TM*). After the final evaluation trust degree being computed, it is the turn of *TM* to manage service trust. *TM* is responsible for three main tasks: ranking the services according to their final evaluation trust degree, recommending the most trustworthy service to requesters and urging the both parties to give the comments on the invocation. Positive or negative comment will play the role of incentives or punishment for services trust, thus can inhibit the network in bad or malicious services. Service rank is also adjusted as the final evaluation trust degree changes with the feedback of requesters who invoke the service.

4.2 Performance Simulation

We simulate our trust evaluate models using NetLogo which is a programmable modeling environment and developed by CCL (Center for Connected Learning) from 1991. In our simulation, the models are tested which are proposed in Section III, i.e, *DlM, eBTM ,ETM* and *FTM*.

We investigate the models from two aspects: effectiveness and efficiency. For effectiveness, the four models are verified whether they can give the correct evaluation to the services. For efficiency, we test the models from the storage and the time needed to accomplish the evaluation.

To measure the effectiveness of the trust models, we employ the invoking success ratio which refers the percentage of the times of successfully invoking services based on trust evaluation to the total times of service invocation. Figure 2 depicts the invoking success ratio goes as the number of execution steps increases. We compare the results of our models (labeled with *WAM*, *MLE*, *BA* and *FIM* in the figure, respectively) with that of the service computing system without trust management (labeled with *none* in the figure).

From the figure, we observe that the invoking success ratio is enhanced greatly using our trust models to manage services. This means trust management with our models can help requesters to invoke the trustworthy services successfully and avoid invoking untrustworthy or malicious services.

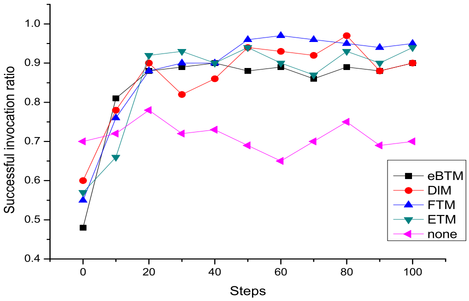


Fig. 2. Effectiveness of trust models

Figure 3 illustrates the time needed for trust computation using the models. The figure shows that the time DIM needs to compute the trust is much less than and much more stable than other three models. This demostrates *DlM* has good scalibility. Figure 4 describes the storage needed to save the direct service trust degree, recommdated servcie trust degree, recommdated trust degree and indexing the services using trust degree. From the figure we find out that the storage increases as the omputation steps increases using all the four models. However, the growth rates of the four models are different with the growth of computational steps. The storage *DlM* needs to save the degree is much less that of the other three models.

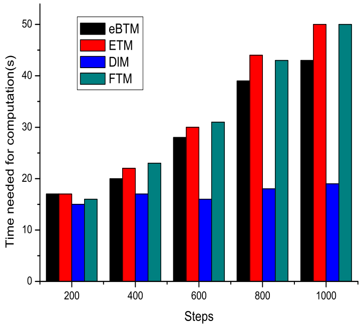


Fig. 3. Time needed for evaluation computation

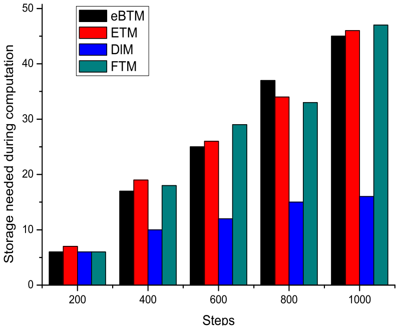


Fig. 4. Storage needed to complete the computation

From the simulation results, we know that that our proposed models can evaluate the trust of service in an effective and efficiently way. For some specific situations, it a good choice to use *DlM* to evluate service trust, such as in wireless computing enviroment.

5 Conclusion

Service computing is now a very important paradigm of distributed computing. Invoking trustworthy service is every service requester respected. In fact, not all services are trustworthy. How to find efficiently the trustworthy services among a large number services, especially on the Internet, is a critical issue to make service computing more applicable. This work proposes a prototype to evaluate and manage service trust in the network. In the prototype, the evaluation of service trust is based on multiple models. Our simulation results show that invoking trustworthy services with our models is much more possible than invoking service without using our models.

Acknowledgement. This work is support by Municipal Nature Science Foundation of Shanghai under Grant No.10ZR1421100 and Foundation of Core Course Construction of USST under grand No. 16000206.

References

1. Maximilien, E.M., Singh, M.P.: Toward Autonomic Web Services Trust and Selection. In: Proceedings of 2nd International Conference on Service Oriented Computing (IC-SOC 2004), New York (November 2004)
2. Olmedilla, D., Lara, R., Polleres, A., Lausen, H.: Trust Negotiation for Semantic Web Services. In: 1st International Workshop on Semantic Web Services and Web Process Composition in Conjunction with the 2004 IEEE International Conference on Web Services, San Diego, California, USA (July 2004)
3. Golbeck, J., Hendler, J.: Inferring trust relationships in web-based social net-works. ACM Transactions on Internet Technology (2006)
4. Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., Sheng, Q.: Quality driven web services composition. In: International World Wide Web Conference (2003)
5. Yu, B., Singh, M.P.: A social mechanism of reputation management in electronic communities. In: Proceedings of Fourth International Workshop on Cooperative Information Agents (2000)
6. Sabater, J., Sierra, C.: Regret: Reputation in gregarious societies. ACM SIGecom Exchanges 3 (2002)
7. Liang, Z.Q., Shi, W.S.: PET: A Personalized trust model with reputation and risk evaluation for P2P resource sharing. In: Proceedings of the 38th Annual Hawii International Conference on System Sciences (2005)
8. Audun, J., Roslan, I., Colin, B.: A survey of trust and reputation systems for online service provision. Decision Support Systems 43(2), 618–644 (2007)
9. Audun, J.: A logic for uncertain probabilities. International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems 9(3), 279–311 (2001)
10. Sepandar, D.K., Mario, T.S., Hector, G.M.: The EigenTrust algorithm for reputation management in P2P networks. In: Proceedings of the 12th International Conference on World Wide Web, Budapest, Hunary, pp. 640–651 (2003)
11. Schmidt, S., Steele, B., Dillon, T.S., Chang, E.: Fuzzy trust evaluation and credibility development in multi-agent systems. Applied Soft Computing Journal 7(2), 492–505 (2007)

Web Application Security Based on Trusted Network Connection[*]

Yongwei Fu and Xinguang Peng[**]

College of Computer Science and Technology
Taiyuan University of Technology
Taiyuan, Shanxi, China
{fuyongwei,pengxinguang}@tyut.edu.cn

Abstract. This paper introduces the security of Trusted Network Connection into Web applications. To solve the security of Web applications and the application limitations of Trusted Network Connection, which is only widely used in LAN and VPN, a new method used for Web application security is presented in the paper based on the thought of Trusted Network Connection. Through the model design and the system realization, it can prove that the thought of Trusted Network Connection can be applied to Web applications and improve the security of Web applications. At the mean time, the thought of Trusted Network Connection can reduce the attack of viruses and trojans and broaden the fields of Trusted Network Connection application.

Keywords: Trusted Network Connection, Web, Viruses, Trojans, Security.

1 Introduction

Trusted Network Connection (TNC) is a network access control technology, which is presented in recent years[1]. The thought of TNC is to provide security guarantee and authentication for the access network terminals, authenticate the security and integrity of the access requestor (AR) by the policy server collecting and evaluating the security state and the credible information of AR[1], execute the access policy on the basis of the AR security level and access authority decision, and finally achieve the access control mechanism of network, prevent the unsafe AR accessing and destroying network. The TNC standard is open, and can be applied to the different brand product. Because it is open, it can be widely used in all kinds of network, provide the integrity protection for the AR and protect users from the intrusion of malicious codes. Currently, TNC is only widely used in LAN and VPN[2], therefore, it is necessary to broaden the TNC application fields. Web application has become the new platform of creating inter-operational distributed applications. For instance, e-bank, e-commerce, egovenment and remote access the intranet. The malware can be able to bypass the security mechanism on the AR and forge the user identity outside

[*] Supported by the Natural Science Foundation of Shanxi Province under Grant No. 2009011022-2; supported by Shanxi Scholarship Council of China Grant No. 2009-28.
[**] Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 420–427, 2011.

to modify personal information in external network. How to enhance its security is also becoming the hotspot and key technology of developing Web application. From the present situation, to integrate easily TNC thought and let users accept, Web application must integrate existing TNC differential mechanism and complete TNC completeness inspection. The testing results are regarded as a visiting basis to determine specific areas of Web application. This paper presents the architecture of Web application that based on TNC thought, and extends the advantages of TNC in LAN and VPN to more extensive network environment. To some extent this architecture can reduce the harm brought by viruses, trojans and malicious codes to protect the security of the access Web application users[3].

2 TNC Architecture

TNC is a department of trusted computing group (TCG) and is also an open standard network access control architecture. TNC is built on trusted computing technology based on host and the main purpose is provided the terminal technology and realized network access control by using credible host. TNC architecture is as shown in Fig. 1.[1].

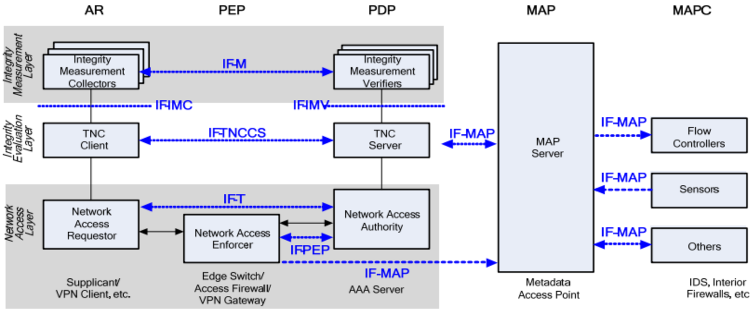


Fig. 1. TNC architecture

Fig. 1. completely describes the latest TNC architecture. TNC architecture is an open network security solutions combined with the existing network access control mechanism. For instance, 802.1X[4] etc. It uses Server-Client model and mainly includes five entities: access requestor (AR), policy enforcement point [5] (PEP), policy decision point[6] (PDP), metadata access point(MAP) and flow controllers and sensors[7] . Where AR is the entity for requesting access network, PDP is the entity for determining network access control policy, PEP is the entity for implementing network access control policy presented by PDP, MAP is an independent metadata server that the user unified centralized stores all kinds of network terminal security state information and policy information, and then constitutes network security information exchange platform, flow controllers and sensors are an control policy that can real-time submit the dynamic security information of terminal to MAP and dynamically adjust the network access behavior on the basis of the security policy information in MAP. Please refer to reference for detailed TNC architecture[1].

An explanation of the process of TNC integrity authentication and access policy authorization is given as follows. First, AR requests to access network and builds a network connection between them. Second, begin the process of integrity authentication, integrity measurement collectors (IMC) starts to collect the integrity information[8], including anti-virus information, firewall information, system update information, etc. Third, AR transmits the integrity information to PDP by PEP, PDP transmits the information to integrity measurement verifiers(IMV) and then IMV verifies the integrity and evaluates security and gives the appropriate network access control policy or remedial measures[9]. Finally, PDP transmits the appropriate network access control policy to PEP executed and transmits remedial measures to AR executed. AR can access network resources until AR meets the integrity verification[1].

The goal of TNC is to make sure the security of AR and put the security extend to the whole network by trusted integrity authentication. As the industry standard, TNC architecture is not completely different from the traditional design, just integrates the existing network access control mechanism and platform framework, so it is not necessary to design interface protocols again for each entity interaction. To some extent this architecture is widely supported and promoted by more and more IT famous enterprises, including Symantec, IBM, Oracle, etc. Some of these enterprises have developed a network security equipment supported TNC[1].

3 Security Analysis of Web Application

The Web application system is a combination of the browser, the Web server and the database. Fig. 2. is the structure of the system. As it can be seen from Fig. 2., the security of the Web application system is applied to the server operating system, the database management system, Web servers, and Web applications, etc.

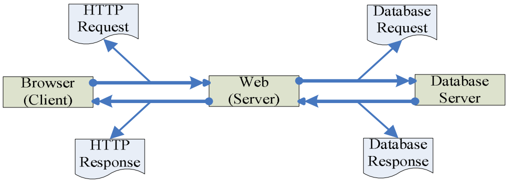


Fig. 2. Structure of web application system

The security of Web application includes two aspects[12]. On the one hand, ensure that the system runs normally and avoid all kinds of unintentional mistakes and damages. On the other hand, prevent the system and data from being used or destroyed by unauthorized users. It is more difficult to guarantee the security of the system in an open network environment. The Web application security architecture is divided into the network system, the operating system, the Web server and the database. The architecture is as shown in the Fig.3.[10].

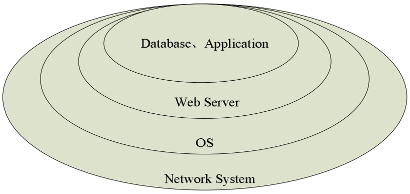


Fig. 3. Web application security system

As it can be seen from Fig. 3., there are five kinds of security bugs[13].

1) Network system security bug. Treats from the network system security are mainly the DDoS attack, the unauthorized remote intrusion, the illegal scanning, the remote detection and the illegal using network resources by network trojan.

2) Operating system security bug. Any operating system has some different degree of loopholes. Especially the default installation and design system has mort serious threats. In addition, it can also reduce the security of systems that installing many operating systems in a server.

3) Web server security bug. The security requirement of Web server usually has two aspects: one is that maintain the integrity of Web contents and the other is that prevent the host be the springboard of intruding network.

4) Database and application program security bug. There are two security problems in the database and application program of Web application. One is the own bug of database, the other is the application program bug of interpreting a script, for instance, the user password being transmitted in the clear, the own bug of Web application program, etc.

4 Web Security Model Based on TNC

Through the analysis of the TNC architecture, we find that original TNC model can not be applied to Web application. The reason is that first, there is not the entity of PEP to deal with all access requests on the internet, for instance, switches, etc; second, there are many online service resources can be accessed by users on the internet, for instance, e-bank, e-commerce, e-government, etc; third, users may access many the entities at the mean time. Because of the differences of network environment, this paper makes some proper expansions of TNC architecture mainly from the following aspects.

Policy server-TNC completeness inspection shows the detailed information of the security protection software and the operating system installed on the host. A poor security policy sever can make the host easily become the attacked target, for instance, a host that not install the latest operating system patches. In Web environment, any services provided access links request completeness inspection to policy server, so, it is very important to prohibit users from using the malicious policy server for completeness inspection.

Information protection-The detailed information of the host in the process of the policy server completeness inspection are exposed to an open network environment and exposes personal secret information to some extent. However, when users access the internet which is lack of network access control mechanism, they do not want to expose any detailed information, so, it is also very important to using integrated information protection mechanism.

Usability and experience of internet users-Compared with professional users, it is difficult for ordinary users to reach the professional level, so, it is very important for users to ensure that TNC completeness inspection process is simple and transparent. In order to meet the current requirements of internet users, initial TNC architecture must be modified. It is also very difficult to meet them totally, so, this paper presents a proper method. Some models designed by this paper and the comparison between them are as follows:

1) Direct validation model-Fig. 4. is this model. The model includes two parts: one is AR and the other is SP. SP executes completeness inspection except providing some common services, for instance, e-bank, e-commerce, e-government, etc. Before authorized AR to access network, SP will provide a completeness inspection to AR. It may lead to expose the detailed information, because while AR transmits its detailed information to SP, SP can learn the identity of users. What is more, when SP makes a decision, it is necessary to update the anti-virus software of AR. Because there are a large number of AR, the workload of task is very big. What is more, it is not also the responsibilities of SP. So this model is not practical. In future, this task may be executed by third-party, and third-party only has one task that is executing completeness inspection of AR.

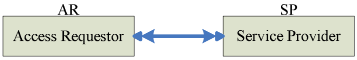


Fig. 4. Direct validation model

2) Relay validation model-Authentication service provider (ASP) provides completeness inspection to AR in this model. Fig. 5. is this model. The TNC message is submitted to ASP by SP, and SP is viewed as transparent agent. However, this method may cause some problems to some extent. Because all TNC messages must be decrypted and encrypted before transmission, this may lead to the bottleneck of SP. Theoretically, SP does not have to understand TNC messages and only needs to transfer forwards. But, AR can not prohibit SP from wiretapping these messages. So, SP can also confirm the detailed information of the security software and system configuration of AR, this will cause the performance information exposed. In addition, if SP can execute TNC completeness inspection on AR, the malicious SP will be able to the policy of ASP and then launch malicious attacks.

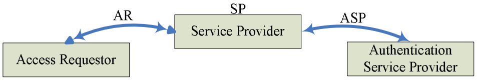


Fig. 5. Relay validation model

3) Annular validation model-Fig. 6. is this model. The difference from above two models is that this model does not have to ASP and AR can communicate directly without SP. The main thought of the model is that SP declares its security requirements taken the form of policy, transmits the policy to ASP by ciphertext, then ASP executes completeness inspection and determines whether AR obeys the policy. Because the messages of completeness inspection are only communicated between AR and ASP, and are not visible for SP, so SP must trust that ASP can execute completeness inspection correctly. This method can reduce the risk of private information exposed. In addition, ASP does completeness inspection on AR for SP in direct validation model, SP only needs a confirmation message from ASP, and then allows AR to access specific resources.

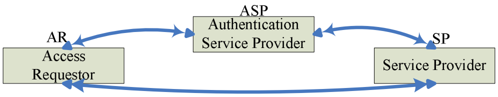


Fig. 6. Annular validation model

5 Implementation Plan

As can be seen from the above three models, it is easy to find that annular validation model is more suitable to be used for TNC completeness inspection. The implementation plan presented by this paper is based on the existing open standards. In order to guarantee the integrity messages transmitted during the process of inspection are effective, the integrity messages must have the digital signature of AR. The message flow based on annular validation model is as shown in Fig. 7.

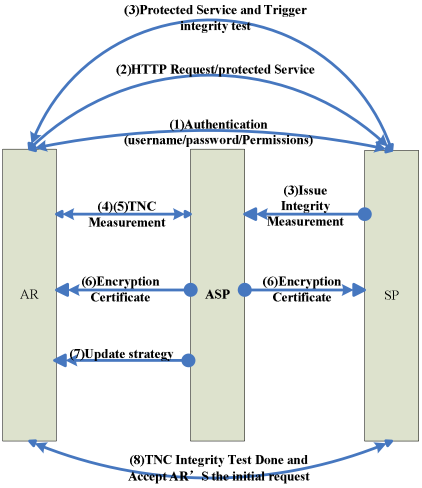

(3)Protected Service and Trigger integrity test
(2)HTTP Request/protected Service
(1)Authentication (username/password/Permissions)
(4)(5)TNC Measurement
(3)Issue Integrity Measurement
(6)Encryption Certificate
(6)Encryption Certificate
(7)Update strategy
AR
ASP
SP
(8)TNC Integrity Test Done and Accept AR' S the initial request

Fig. 7. Message flow model

The specific procedures are as follows:

1) Users access the Web application provided by SP through Secure Hypertext Transfer Protocol (HTTPS) and are identified through user name and password.

2) When users access the limited Web services, users will be inspected through TNC completeness inspection. This can reduce the probability that malicious codes attack the online services.

3) When SP executes TNC completeness inspection, the related information and access policy are transmitted to policy server in ciphertext. Only policy server can access these messages. This can protect them from not redirected to malicious software.

4) The plug-in installed in Brower executes the request of TNC completeness inspection and transmits the result to policy server by digital signature. The information between AR and policy server can use a standard TNC TNCCS message[11]. Because TNCCS message is based on XML, the information of HTTP is transmitted by SOAP.

5) After policy server receives the cryptographic integrity information, policy server executes TNC completeness inspection on the basis of TNC standard[8-9].

6) If the integrity information of AR meets the policy provided by SP, policy server will transmit two corresponding cryptographic certificates to AR and SP accessed respectively. These messages are included in the TNCCS message and are readable.

7) If not, policy server will provide matching update policy to AR. And this is also transmitted in TNCCS way.

8) SP deciphers the message received from policy server and then executes corresponding access policy, which is that permit or reject the AR request of step two.

A TNC completeness inspection is completed and then the authorized AR can communicate datum to SP. In the process of completeness inspection and access, if AR does not obey the policy, TNC completeness inspection will be executed once more, even if AR is legal, for instance, the server having user name and password. Similarly, if the certificate has been expired, TNC completeness inspection will be also executed once more.

Because only policy server certifies the access rights of AR, users can not fake it. Thus, this plan is secure and feasible.

6 Conclusions

This paper discusses TNC architecture and application, and then discusses Web application and security holes. On the basis of muttons, this paper presents a kind of Web application based on TNC thought and a kind of annular validation model. This method can protect private information and is easy to achieve. This model considers TNC completeness inspection as a part of the process that Web application identifies AR. Our future work focuses mainly on estimating the performance of this model.

Acknowledgments. This work is supported by the Natural Science Foundation of Shanxi Province under Grant No. 2009011022-2; supported by Shanxi Scholarship Council of China Grant No. 2009-28.

References

1. TCG Trusted Network Connect TNC Architecture for Interoperability Specification Version 1.4 Revision 4, Published (May 18, 2009)
2. Liu, H., Wei, G.: Application of trusted computing compliance in VPN. Computer Applications 26(12), 2935–2937 (2006)
3. Garfinkel, S., Spanfford, G.: Web Security, Privacy & Commerce. China Machine Press, Beijing (2004)
4. IEEE802. Port-based network access control. IEEE Std 802.1X -2001 (June 2001)
5. TCG Trusted Network Connect TNC IF-PEP: Protocol Bindings for RADIUS Specification Version 1.1 Revision 0.75, Published (February 2007)
6. Christine, M.T.: Trusted Network Connect TNC Architecture for Interoperability Specification Version 1.2 Revision4 (EB/OL) (April 15, 2008) (December 19, 2008)
7. TCG Trusted Network Connect TNC IF-MAP Metadata for Network Security Specification Version 1.0 Revision 25, Published (September 13, 2010)
8. TCG Trusted Network Connect TNC IF-IMC Specification Version 1.2 Revision 8, Published (February 5, 2007)
9. TCG Trusted Network Connect TNC IF-IMV Specification Version 1.2 Revision 8, Published (February 5, 2007)
10. Leveson, N.G.: Safeware: System Safety and Computers – A Guide to Preventing Accidents and Losses Caused by Technology (1995)
11. TCG Trusted Network Connect TNC IF-TNCCS: TLV Binding Specification Version 2.0 Revision 16, Published (January 22, 2010)
12. Kaufman, C., Perlman, R., Speciner, M.: Network Security—Private Communication in a Public World. Publishing House of Electronics Industry, Bingjing (2004)
13. Garfinkel, S., Spanfford, G.: Web Security, Privacy & Commerce. China Machine Press, Beijing (2004)

Model Checking for Asynchronous Web Service Composition Based on XYZ/ADL

Guangquan Zhang[1,2], Huijuan Shi[1], Mei Rong[3], and Haojun Di[1]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
[2] State Key Laboratory of Computer Science, Chinese Academy of Science, Beijing, China
[3] Shenzhen Tourism College, Jinan University, Shenzhen, China
gqzhang@suda.edu.cn

Abstract. Concerned with Web service composition, this paper proposes a model checking method of verifying asynchronous communication behaviors and timed properties. Firstly, analyzing Web service composition from software architecture, the interactive behaviors and timed properties are described by XYZ/ADL based on temporal logic language. Secondly, timed asynchronous communication model (TACM) which accords with the specification of model checker UPPAAL is proposed. Finally, based on the transition from XYZ/RE communication commands to TACM, the correctness of asynchronous communication behaviors of the service composition system can be verified by UPPAAL.

Keywords: Web service composition, XYZ/ADL, asynchronous communication, timed properties, model checking.

1 Introduction

Service Oriented Architecture (SOA), which supports reusability, loose coupling, interoperability, platform independent, is an innovative architecture paradigm. Web service technology, a widely used distributed computing technology, has become main implement way of SOA. With the fast development of e-business, single Web service couldn't meet the complicated demand, which causes service composition to be a research hotspot of software service field.

Web service composition implements big service function through the communication and coordination among small Web services. The communication includes synchronous and asynchronous communication. The existing works [1-3] almost are based on synchronous communication. However, the nature of distributed systems and particularly Web services are asynchronous, which makes these approaches restrictive in real application scenarios. To overcome such limitations, some works tried to consider the asynchronous communication. Ref. [4] compared the difference and similarity between synchronous and asynchronous communication of Web service in detail, and presented up-bottom and bottom-up asynchronous interaction model of it. Ref. [5] researched asynchronous communication using recall, and proposed a reliability interaction pattern of Web service. Ref. [6] discussed choreography, orchestration and session of Web service by asynchronous π-calculus. However, these works didn't

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 428–435, 2011.

consider timed properties when analyzing the asynchronous communication. Additionally, we verify timed properties of composite Web service in Ref. [7], but it didn't support the asynchronous communication.

To resolve above problems, we propose a model checking method to verify the interactive behavior and timed properties of Web service. Firstly, use XYZ/ADL based on temporal logic language to precisely characterize the interactive behaviors and timed properties of Web service. Then propose a Timed Asynchronous Communication Model (TACM), which can be verified directly using model checker UPPAAL. By translating the XYZ/RE communication commands to TACM, the related properties of Web service can be verified by UPPAAL. Compared to other approaches, it need not convert the system model to the specification of model checker, which simplifies the verification process.

2 Web Service Composition Based on XYZ/ADL

Compared Web service composition and Software Architecture (SA), we can find that they have some similarities. Web service can be considered as component. The interactive rules of Web services correspond to connector. The whole layout of Web service composition can be regarded as configuration. These relations provide support and foundation for researching Web service composition from a higher level. It is convenient to realize formal verification using XYZ/ADL, an architecture description language (ADL) based on temporal logic, to describe Web service composition.

2.1 XYZ/E

XYZ/E is an executable temporal logic language, in which basic unit is conditional element. There are two forms:

$$\mathtt{LB=y \wedge R \Rightarrow \$Ov=e \wedge \$OLB=z} \quad (1)$$

$$\mathtt{LB=y \wedge R \Rightarrow @(Q \wedge LB=z)} \quad (2)$$

Where, R and Q are first-order logic formula, R represents the condition part, Q and \$Ov=e are action parts, LB is control variable, y and z are the entry and exit label respectively, => represents logic implication. In formula (1), conditional element defines the transition relation of the adjacent states. The symbol @ in formula (2) can be next operator \$O or final operator < >, they represent the abstract specification of a program.

2.2 XYZ/ADL and XYZ/RE

XYZ/ADL, supports the concept of component, connector, configuration in SA, is an ADL extended from XYZ/E. It can describe system from formal specification to executable program under unified logic framework. A completed XYZ/ADL description is shown as follows.

```
%COMPONENT COM==[
%PORT P:Record (%CHN A:DATATYPE1; %CHN B: DATATYPE2)
```

```
  %PROPERTY== […]
  %COMPUTATION== […]]
  %CONNECTOR CON== [
  %ROLE SourceData==OUT(DT1,v1);
  %ROLE  SinkData==IN(DT2,v2) ;
  %GLUE== […]]
  %ATTACHMENTS==[ComIns.Port#ConIns.Role;ComIns.Port##Por
t;…]
```

Where, the internal specification of component uses a XYZ/E unit to represent the behavior of different layer of Web service. If involves timed restraint it can be described by XYZ/RE. Extended the temporal operator $O, < >, [ ], $U, $W in XYZ/E in order to make them capable to express the real time lower limit(l) and upper limit(u), so we get real time XYZ/E, namely XYZ/RE [8]. XYZ/RE conditional element has tow basic forms:

```
  LB=L0∧R⇒$O{l,u}(Q∧LB=L1)
  LB=L0∧R⇒@{l,u}(Q∧LB= L1)
```

3 Asynchronous Communication of Web Services

Asynchronous communication is important for building robust Web services [4]. This section presents a survey of related work on analyzing and modeling the asynchronous Web service.

3.1 Analysis of Asynchronous Communication

Web services interact by exchanging messages, which includes sending and receiving message. They are denoted by !m and ?m respectively. The real Web service composition often involves timed properties, which are crucial properties of service interaction [9]. When modeling Web services, we use the standard timed automata clocks to capture the timed properties. The values of these clocks increase with the passing of time. Transitions are labeled by timed constraints, called guards, and resets of clocks. Figure 5 shows an example of timed asynchronous services interaction. To assure the correct interaction between asynchronous services, each service is equipped with a queue to store the incoming message. We assume the queue is unbounded and messages can be consumed in any arbitrary order.

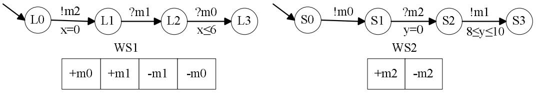


Fig. 1. A simple example of timed asynchronous service

Let us not consider the timed constraint at first. The service WS1 starts by sending message m2, which is stored (indicated by '+') in the queue of WS2. On the other hand, WS2 can send message m0, which is added in the queue of WS1. The service WS1 remains blocked because the message m1 can't be available. Then WS2 consumes (indicated by '-') message m2 and send message m1. After that, WS1 can consume message m1 and m2. Consequently, WS1 and WS2 finish interaction successfully and both queues become empty. If consider the timed constraints, the communication process is shown as follows. WS1 sends message m2 and reset the clock (x=0).WS2 reset the clock (y=0) after consuming message m2. Then WS2 sends message m1 after 8 and within 10 units of time ($8\leq y\leq 10$). Finally, WS1 must consume message m0 within 6 units of time. However, m0 can only be consumed after consuming m1, i.e., after 8 units of time. Obviously, WS1 represents a timed conflict at state L2 and can't transfer to final state L3.

3.2 Timed Asynchronous Communication Model

To model the asynchronous Web service, we first abstract clocks and messages to two variables of distinct types respectively. Initially, the values of the variables equal to zero and the message queues are empty. The clock variables are continuous. Their values increase automatically with the passing of time and can be reset to zero. The message variables are discrete and their values can only equal to 0 or 1.When send message, this message is added into the queue and the value of associated variable is set to 1.When receive message, first check if this message be in the queue or the value of related variable equals to 1. If the result is true, consume this message and set the value of corresponding variable to 0. According to above analysis, Web service asynchronous communication model is defined as follows:

Definition 1. (Timed Asynchronous Communication Model) A TACM is a tuple (S, s_0, F, C, M, A, T) such that S is a set of states, s_0 is initial state ($s_0\in$ S), F is the set of final states ($F\subseteq S$), C is the set of clock variables, M is the set of message variables, $A:M\rightarrow \{?,!\}$ is a labeled function, it assigns an action of receive (?) or send (!). $T\subseteq S\times g(C, M)\times u(C, M)\times S$ is a set of state transition. A transition from state s to state s', denoted (s, g, u, s'), will be triggered if the guard condition g is satisfied, and it will execute action u to update the values of clock and message variables.

The set of constraints over C and M, denoted g(C, M), is defined as follows:

$g(C,M)=$ true $\mid x \sim v \mid y == 1 \mid \psi 1 \wedge \psi 2$, where$\sim\in\{\leq,<,==,!=,>,\geq\}$,$x\in C$, $y\in M$,$\psi 1$,$\psi 2\in g(C,M)$, $v\in \mathbb{R}_{\geq 0}$is a positive real constant.

The set of actions which update C and M, denoted u(C, M), is defined as follows:

$u(C,M) = x{:=}n \mid y{:=}0 \mid y{:=}1 \mid \varphi 1\wedge\varphi 2$, where $x\in C$, $y\in M$, n is a constant, $\varphi 1,\varphi 2\in g(C,M)$.

If A (m) =? then the constraints over M in g(C,M) is m==1, the updating action over M in u(C,M) is m:=0; If A (m) =! then there is no constraints over M, the updating action over M in u(C,M) is m:=1.

TACM is a behavioral model, which can precisely characterize the asynchronous interactive behaviors of Web service and related timed restraints. The elements describing behavior consist of finite states, conditions of triggering transition and actions executed during transition, which completely correspond to the elements of XYZ/E conditional element. Additionally, TACM accords with the specification of model checker UPPAL.

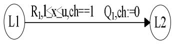
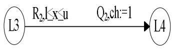
4 Model Checking of Asynchronous Web Service Composition

Model checking [10] is a method for formally verifying finite-state concurrent system. Specifications about system are expressed as temporal logic formula, and test automatically the given model of system whether meets the specifications. At present, a lot of distinctive model checking tools have been used widely, such as SPIN, NuSMV, UPPAAL and so on. This paper selects UPPAAL to verify Web service composition. The main reasons are that UPPAAL can verify the timed properties of Web service and TACM satisfies the input model of UPPAAL.

The specific verification process of asynchronous Web service composition based on XYZ/ADL is shown as follows. Firstly, use XYZ/ADL to describe the Web service composition. Then translate the XYZ/RE communication commands in XYZ/ADL to TACM. Finally, express the specification with a CTL formula, and put both TACM and the specification into UPPAAL to verify the relative properties.

In the XYZ/ADL description of Web service composition, the interactive behaviors are described by XYZ/E unit and the related timed properties are represented by XYZ/RE. Web services interact by exchanging messages. Obviously, the mostly used statements are input and output sentences, namely the communication commands. Therefore, we only realize the mapping from communication commands to TACM. The table1 shows the mapping rule.

Table 1. Mapping rule from communication commands to TACM

	XYZ/RE	TACM
Input sentence	LB=L1 $\wedge$ R_1 $\wedge$ ch?$\Rightarrow$\$O{l,u}(Q_1 $\wedge$ \$OLB=L2)	
Output sentence	LB=L3 $\wedge$ R_2 $\wedge$ ch!$\Rightarrow$\$O{l,u}(Q_2 $\wedge$ \$OLB=L4)	

5 Case Analysis

Let us present a composite service BuyBook to illustrate our approach. It involves three Web services: Customer, BookShop and Bank. The interactive process and timed constraints are briefly summarized as follows. Customer sends a *searchBook* request to BookShop. After searching, if there is stock BookShop returns *bookID* and *bookPrice* successively, or returns *outStock* with in 5 units of time. Customer receives the message *bookPrice* at first, sends Bank a request *balanceInquire* within 2 units of time to search the balance of account, and then receives *bookID*. Bank returns the message *balance* within 3-5 units of time. If the balance is enough Customer sends a message *payInfo* to Bank, or sends *cancel* to BookShop within 5-10 units of time. After paying the bill, Bank sends a message *payConfirm* within 3 units of time to inform BookShop that Customer has paid the bill. Finally, BookShop sends *invoice* to Customer within 4 units of time. This is an example of timed asynchronous communication. The XYZ/ADL description codes are shown as follows.

```
%COMPONENT Customer==[
LB=S0∧!searchBook⇒$OLB=S1;
LB=S1∧?outStock⇒$OLB=S2;
LB=S1∧?bookPrice⇒$O(x=0∧LB=S3);
LB=S3⇒$O{0,2}(!balanceInquire∧LB=S4);
LB=S4∧?bookID⇒$OLB=S5;
LB=S5∧?balance⇒$O(x=0∧LB=S6);
LB=S6∧!cancel⇒$O{5,10}(LB=S7);
LB=S6∧!payInfo⇒$O{5,10}(x=0∧LB=S8);
LB=S8∧?invoice⇒$OLB=S9;]
%COMPONENT BookShop==[
LB=L0∧?searchBook⇒$O(y=0∧LB=L1);
LB=L1∧!outStock⇒$O{0,5}(LB=L2);
LB=L1∧!bookID⇒$O{0,5}(LB=L3);
LB=L3∧!bookPrice⇒$OLB=L4;
LB=L4∧?cancel⇒$OLB=L5;
LB=L4∧?payConfirm⇒$O(y=0∧LB=L6);
LB=L6⇒$O{0,4}(!invoice∧LB=L7);]
%COMPONENT Bank==[
LB=M0∧?balanceInquire⇒$O(z=0∧LB=M1);
LB=M1⇒$O{3,5}(!balance∧LB=M2);
LB=M2∧?payInfo⇒$O(z=0∧LB=M3);
LB=M3⇒$O{0,3}(!payConfirm∧LB=M4);]
```

According to the mapping rule from XYZ/RE communication commands to TACM, we can obtain the TACM of Customer, BookShop and Bank. Figure 6 shows the models, where S2, S7 and S9 in Customer are final states. L2, L5 and L7 are final states of BookShop. In Bank, M2 and M4 are final states. These final states are all labeled in green.

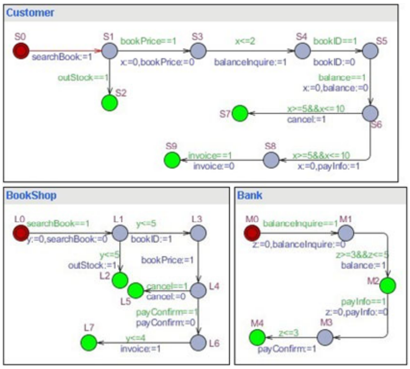


Fig. 2. TACM of Customer, BookShop and Bank

We use model checker UPPAAL to mainly verify deadlock, safety properties and liveness properties.

(1) Deadlock indicates that Web service stay at some state, which doesn't satisfy transition condition, and can't continue to interact. In TACM, no deadlock is equivalent to check if all the services reach their final states. In other words, when the services reach their final states, all the values of the message variables must be equal to zero. It is specified as the following CTL formulas:

E<>(Customer.S2 and BookShop.L2) or (Customer.S7 and BookShop.L5 and Bank.M2) or (Customer.S9 and BookShop.L7 and Bank.M4) and (*searchBook*==0 and *bookPrice*==0 and *balanceInquire*==0 and *bookID*==0 and *balance*==0 and *payInfo*==0 and *invoice*==0 and *cancel*==0 and *outStock*==0 and *payConfirm*==0).

(2) Safety properties are on the form: "something bad will never happen". In this example, the user requests "Customer receive the message *invoice* without paying the bill" never happen. It is specified as the following CTL formulas:

A[] not (Customer.S7 and Customer.S9).

(3) Liveness properties are on the form: "something will eventually happen". In this example, the user requests "Customer receive *invoice* within 10 units of time after paying the bill" eventually happen. It is specified as the following CTL formulas:

A<>Customer.S8 imply Customer.S9 and x<10.

6 Conclusion

Web service composition is one of the research hotspots in Service Oriented Computing (SOC). Asynchronous communication is an important feature of information exchange in Web service composition. However, the current works seldom consider timed properties and asynchronous communication. Therefore, we propose a timed asynchronous communication model TACM, which can precisely describe the timed properties and asynchronous communication behaviors. Additionally, TACM can be directly put into UPPAAL to verify the related properties. On the other hand, we analyze Web service composition from software architecture, which makes it possible to control the layout of the whole system. As future work, we will improve TACM to apply to more asynchronous communication scenarios. What's more, we will consider how to capture and handle all kinds of exceptions occurring in the interactive process of Web service.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China under Grant No.60973149; the Natural Science Foundation of Jiangsu Province, China; the Open Funds of State Key Laboratory of Computer Science of Chinese Academy of Science under Grant No.SYSKF0908; the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant No.08KJB520010.

References

1. Pistore, M., Roveri, M., Busetta, P.: Requirements-driven verification of Web services. In: WSFM 2004, pp. 95–108 (2004)
2. Lei, L.H., Duan, Z.H.: An extended deterministic finite automata based method for the verification of composite Web services. Journal of Software 18(12), 2980–2990 (2007) (in Chinese)

3. Foster, H., Uchitel, S., Magee, J., Kramer, J.: Model-based verification of Web service compositions. In: IEEE ASE 2003, pp. 152–161 (2003)
4. Fu, X., Bultan, T., Su, J.: Synchronizability of conversations among Web services. IEEE Transactions on Software Engineering 31(12), 1042–1055 (2005)
5. Pallickara, S., Fox, G., Pallickara, S.L.: An analysis of reliable delivery specifications for Web services. In: 10th International Conference on Information Technology: Coding and Computing, pp. 360–365 (2005)
6. Vieira, H.T., Caires, L., Seco, J.C.: The conversation calculus: A model of service oriented computation. In: 17th European Symposium on Programming, pp. 269–283 (2008)
7. Zhang, G.Q., Rong, M., He, Y.L., Zhu, X.Y., Yan, R.J.: A refinement checking method of Web service composition. In: IEEE SOSE 2010, pp. 103–106 (2010)
8. Tang, Z.S., et al.: Temporal logic programming and software engineering. Science Press, Beijing (2002) (in Chinese)
9. Guermouche, N., Godart, C.: Timed model checking based approach for Web services analysis. In: IEEE ICWS 2009, pp. 213–221 (2009)
10. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. MIT Press, Cambridge (1999)

Specification and Verification of Data and Time in Web Service Composition

Guangquan Zhang[1,2], Haojun Di[1], Mei Rong[3], and Huijuan Shi[1]

[1] School of Computer Science and Technology, Soochow University, Suzhou, China
[2] State Key Laboratory of Computer Science, Chinese Academy of Science, Beijing, China
[3] Shenzhen Tourism College, Jinan University, Shenzhen, China
gqzhang@suda.edu.cn

Abstract. The verification of Web service composition has been widely acknowledged as a challenging problem. In this paper we present a method based on data and time aware service model to validate property of Web service composition. First we translate Web service composition specification to formal model which contains data related information and time related information, and then translate this model to UPPAAL specification, at last the correctness of Web service composition is verified through the UPPAAL tool.

Keywords: Web service composition, Verification, Model checking, UPPAAL.

1 Introduction

With the development of the Internet and electronic commerce technology, Service Oriented Architecture (SOA) is now the trend of software development. As the main way to realize SOA, Web service technology is also undergoing fast development and application. Web service is a self-described and self-contained software module which is based on network, and it has the characteristics of loose coupling and strong independence. As individual Web service has its limitations, through the way of compositing Web services to meet user's requirements has become an inevitable trend, but how to ensure the quality of Web service compositions has also become the focus in academia and industry sphere. Formal verification method is one of the important ways to ensure the correctness of software, and the most common formal verification methods are divided into two categories: deductive verification and model checking.

Among them, model checking is a fully-automatic verification technique for the verification of finite state systems [1]. The main idea of model checking can be concluded as follows: we use state transition system S to model the behavior of a system, and use temporal logic formula F to describe the properties of the system. "Whether the system satisfy the desired properties" can be converted to a math problem which is "whether state transition system S is a model of formula F".

2 Related Work

Currently, there has been a lot of research work on the validation of property of Web service composition by using model checking techniques. Fu et al. [2] present a set of

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 436–443, 2011.

tools and techniques for analyzing interactions of composite Web services which are specified in BPEL, and use SPIN model checker verify the properties of Web service compositions. Diaz et al. [3] present a method of automatic translation of WS-CDL choreographies to timed automata, and analyze the system behavior using UPPAAL tool. Zhao et al. [4] translate WS-CDL to the input language of the model checker SPIN, which allows us to automatically verify the correctness of a given choreography. Guermouche et al. [5] study a model checking based approach that deals with checking the compatibility of the choreography where the Web services support asynchronous timed communications. Peng et al [6] propose a pi-calculus based automatic method to verify the correctness of composite service, and with the importing of the new verification method, the precision of service discovery process has been improved.

We also have done similar work in this respect. In [7], we studied the correctness of Web service compositions and their formal verification method in a high and abstract view, using software architecture description language XYZ/ADL to describe Web service compositions and model checking tool UPPAAL to realize the automatically verify to the properties of Web service compositions.

However, in these studies, the majority just consider the verification of behavior properties of Web service compositions (such as Safety and Liveness), seldom consider the data related information in the progress of interactions of Web services, and lack verification of data related properties. Besides, the validation of timed properties of Web service compositions is also not very comprehensive. Based on this, in this paper we extend finite automaton, and present a model checking method based on data and time aware service model (DTSM), then simulate and verify property of Web service compositions.

3 Data and Time Aware Service Model

In this section we will present the data and time aware service model which can depict data related information and time related information in the progress of interactions of Web services. Now we first introduce an example of Fruit Sales Service (FSS) as the basis of the discussion in this paper.

Figure 1 shows FSS which is composed by InventoryService and BillingService. In this scene, if buyer wants to buy fruits, he can submit queries to FSS, according to queries information, FSS will call InventoryService to return fruitdetails message containing fruit types, price, stocks etc, or return the searchinvalid message; if buyer is ready to buy fruits, he can send a buyorder message. After receiving it, FSS will call BillingService to check. If it is approved, FSS will send an orderconfirm message, otherwise return an orderinvalid message; At last, after sending payment message, buyer will receive the paymentconfirm message, otherwise receive paymentinvalid message. In addition, in the interactive progress, buyer should receive feedback message in a short time after sending message to FSS. FSS will also make the corresponding processing if buyer has no response in a long time.

Scenario above shows the requirements of depicting data related information and time related information. Regarding data related information, in the whole interaction progress, obviously it needs to meet "the fruit information (such as fruit names)

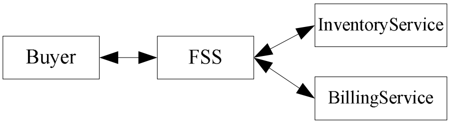


Fig. 1. An example of web service composition

contained in buyorder message sent by buyer and it should be consistent with information contained in payment message" and "only if buyer pay the right amount, paymentconfirm message will be sent by FSS ", So these properties demand to depict hidden data related information such as fruit types, unit price, total price and so on. When it comes to time related information, it needs to meet "it will continue to make effective trade only if buyer receives feedback message sent by FSS within the stipulated time ", so this property demands to depict time related information.

In order to depict data related information, we use the concept of message data to represent data related information. We define the data content contained in messages exchanged between Web services which can be expressed and have practical significance as message data, and abstract messages as the set of message data.

Definition 1. Let M be a set of messages, D be a set of message data variables, V be a set of value of message data, we can define a mapping: L: M×D→V.

For every message m $\in$ M, L $(m,d_i) \in$ V represents message data value vi which is contained in message m .In addition, we can suppose that L uniquely identities every message, that is, it does not exist messages m_1, $m_2 \in$ M, such that L (m_1, d) = L (m_2, d), where $m_1 \neq m_2$, $d \in$ D.

Because that message data contained in message has a variety of data types. To set of message data variable D, and set of data types T, we define a mapping F: D→T, that is, for each message data variable d $\in$ D belongs to data type F (D).

In order to depict time related information, we use clock in timed automata to express time related information. The clock is a nonnegative real number of variable, all the clocks pass in a unified speed. At the start the clock generally set to zero, when event occurs, the clock can be reset to zero. A clock constraint means make the comparison with the clock value and a timed constants. It is defined as follows.

Definition 2. Let X be a set of clock variables, Δ be a set of clock, the grammar of clock variable constraint δ defined as follows:

$$\delta ::= x \leq c | x \geq c | x < c | x > c | \neg \delta | \delta \wedge \delta$$

Among them, clock variable $x \in X$, and c is a nonnegative real number constants.

In this paper, we extend finite automaton by using message data to represent data related information, and using clock in timed automata to represent time related information. The model of web service is defined as follows.

Definition 3. Data and time aware service model (DTSM) is a tuple $\sum$= (S,s_0,F,M,X,D,I,I',R) where:

- S is a finite set of states;
- $s_0 \in S$ is a initial state;

- $F \subseteq S$ is a set of final states;
- M is a set of messages;
- X is a finite set of clocks;
- D is a finite set of message data;
- R is a set of transitions such that $R \subseteq S \times M \times \delta(X) \times \varphi\ (D) \times 2^X \times S$.
- I: $S \rightarrow \Delta(X)$ is a function which assigns timed constraints to the states;
- I':$S \rightarrow \Phi(D)$ is also a function which assigns message data constraints to the states;

In the progress of interactions, $(s,m,\delta,\varphi,\gamma,s')$ represents the transition from state s to state s', m is a message received or sent by state s, δ is a timed constraints over X, φ is a message data constraints over D, they are satisfied when the transition occurs, $\gamma \subseteq X$ is a set of clocks which reset to zero when the transition occurs. In addition, a message data value L (m, d) involving in a transition progress is the data content contained in message m which is received or sent by the current state s.

Definition 4. Composition of data and time aware service model $\sum_1$and $\sum_2$ is$\sum=(S_1 \times S_2, s_{01} \times s_{02}, F_1 \times F_2, M_1 \cup M_2, X_1 \cup X_2, D_1 \cup D_2, I, I', R)$,such that $((s_1,s_2),m,\delta,\varphi,\gamma,(s_1',s_2')) \in R$ iff :

- $(s_1,m,\delta,\varphi,\gamma,s_1') \in R_1$, $m \notin M_2$, $s_2'=s_2$, or
- $(s_2,m,\delta,\varphi,\gamma,s_2') \in R_2$, $m \notin M_1$, $s_1'= s_1$, or
- $(s_1,m,\delta_1,\varphi_1,\gamma_1,s_1') \in R_1,(s_2,m,\delta_2,\varphi_2,\gamma_2,s_2') \in R_2$, $m \in M_1, m \in M_2, \delta=\delta_1 \wedge \delta_2$, $\varphi=\varphi_1 \wedge \varphi, \gamma=\gamma_1 \cup \gamma_2$.

Obviously, through the extending on finite automaton, data and time aware service model (DTSM) has the ability to express the data related information and time related information. In addition, in order to better focus on expressing data related information and time related information, we overlook some details, such as the problem of the communication of messages exchanged between services, the storage of data related information, and state-space explosion caused by data related information.

4 Model Checking of Web Service Compositions

4.1 From BPEL to DTSM

BPEL (Business Process Execution Language) is the de facto standard to describe the interactions of the individual web service in both abstract and executable [6]. In this section we will introduce the conversion rules between BPEL and data and time aware service model (DTSM).

Figure 2 shows the progress of conversion between BPEL and DTSM. < receive > activity can receive a message from external environment. This is a basic activity and it can be accomplished by one transition action; <while> activity belongs to structured activities. Structured activities prescribe the order where a collection of activities occur, and they define multiple activities in internal environment, so in DTSM they will be accomplished by multiple transition actions.

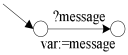
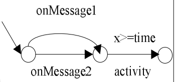
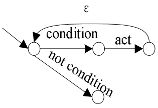
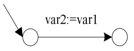
BPEL		DTSM	BPEL		DTSM
receive	<receive operation=”op” variable=”var” />		pick	<pick> onMessage1 onMessage2 <onAlarm> <for>time</for> activity act </onAlarm> </pick>	
while	<while> <condition> condition </condition> activity act </while>		assign	<assign> <copy> <from var=”var1”> <to var=”var2”> </copy> </assign>	

Fig. 2. From BPEL to DTSM

<assign> activity contains the expression of data related information, and has the ability of copy message data, so it can be used to operate data related information. <pick> activity contains the expression of time related information. In DTSM, the expression which contains clock variables x is used to represent timed constraints. In <pick> activity, after an event has been selected, the activity will be executed after certain time. In DTSM, it can be explained that a translation action will be executed only if the clock variables x is greater than certain time constants.

4.2 From DTSM TO UPPAAL Specification

UPPAAL is a tool box for validation and verification of real-time systems. It is based on timed automata, which is a finite automaton with clocks. The main advantage of this tool is the efficiency and practicability. Compared with other model checking tools (such as Spin, NuSMV, etc.), the input language of UPPAAL is the most similar to DTSM, so in this paper we choose UPPAAL to validate the properties of Web service compositions.

Because of the semantic similarity between UPPAAL specification and DTSM, most of elements can be translated directly. But we need to consider the conversion of time related information and data related information. Regarding the conversion of time related information, because UPPAAL is based on timed automata, so it has the ability to express time related information. In addition the explanation of time in DTSM is also based on the concept of clock, so both of them are similar in representing time related information, and there is no need to be converted.

Regarding the conversion of data related information, in DTSM, we use message data to represent data related information, but UPPAAL specification can not afford it. So we can define a conversion function Trans: L (m, d) → Var, where m ∈M is a set of messages, d ∈D is a set of message data variables, Var is s set of integer variables. Thus we can mark message data contained in messages directly by using integer variables.

4.3 Case Study

Taking fruit sales service FSS as an example, We suppose that in the progress of interactions, buyer should receive the feedback message within 2 unit time after he sends the message to FSS, and FSS will give up trading if he does not receive the response from buyer within 10 unit time. Besides, buyer will send buyorder message to FSS only if the price of fruit is less than 10 dollars. When receiving payment message sent by buyer, FSS will send paymentconfirm message only if that the total amounts paid by buyer is right

According to requirements mentioned above, the DTSM model of buyer and FSS can be shown in figure 3.

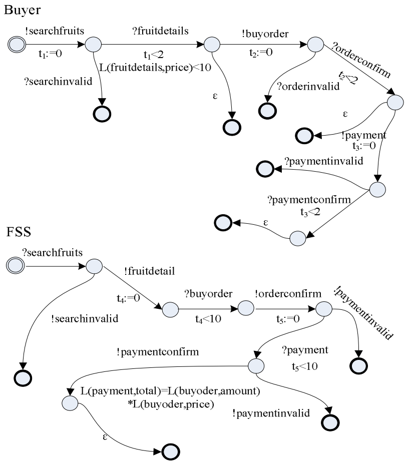


Fig. 3. The DTSM model of buyer and FSS

According to the rules of conversion, we can convert DTSM model of buyer and FSS to UPPAAL specification, and the results can be shown in figure 4.

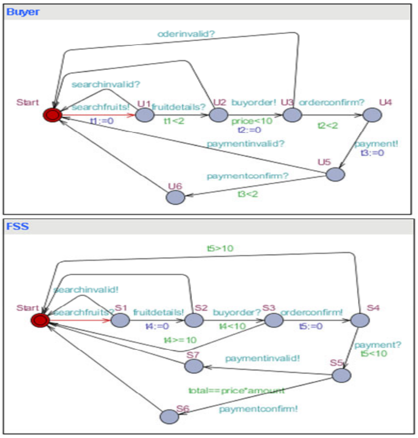
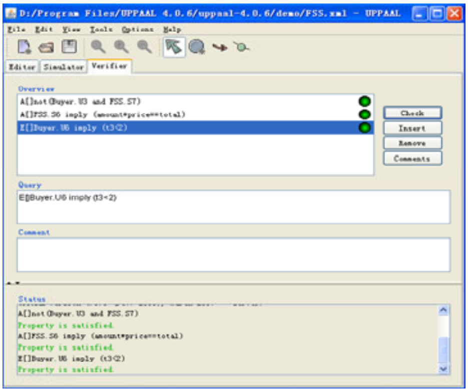


Fig. 4. The UPPAAL specification of buyer and FSS

Now we can use UPPAAL tool to analyze and verify the property of Web service compositions, and the results of verification can be shown in figure 5.


D:/Program Files/UPPAAL 4.0.6/uppaal-4.0.6/demo/FSS.xml - UPPAAL
File Edit View Tools Options Help
Editor
Simulator
Verifier
Overview
Check
Insert
Remove
Comments
Query
Comment
Status

```
A[]not (Buyer.U3 and FSS.S7)
A[]FSS.S6 imply (amount*price==total)
E[]Buyer.U6 imply (t3<2)
```

Query: E[]Buyer.U6 imply (t3<2)

```
A[]not (Buyer.U3 and FSS.S7)
Property is satisfied.
A[]FSS.S6 imply (amount*price==total)
Property is satisfied.
E[]Buyer.U6 imply (t3<2)
Property is satisfied.
```

Fig. 5. The results of verification

1) Behavior property: at any moment, it is not possible that FSS canceled buyer's orders, in case of that buyer has paid the bill. This property can be expressed as:

A[]not (Buyer.U3 and FSS.S7);

2) Data related property: only that buyer pays the right amount of payment, FSS will send the paymentconfirm message. This property can be expressed as:

A[]FSS.S6 imply (amount*price==total);

3) Timed property: after receiving buyorder message sends by buyer, FSS should send orderconfirm message in two minutes. This property can be expressed as:

E[]Buyer.U6 imply (t3<2).

5 Conclusion

In this paper, we have extended the finite automaton, and presented a model checking method based on data and time aware service model. Through examples, we have analyzed and verified the property of Web service compositions. Our future work will focus on further extension of DTSM, in order to enhance the model's expression abilities. Besides, we will also consider using bounded model checking techniques to solve the problem of state-space explosion caused by excessive message data.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China under Grant No.60973149; the Natural Science Foundation of Jiangsu Province, China; the Open Funds of State Key Laboratory of Computer Science of Chinese Academy of Science under Grant No.SYSKF0908; the Natural Science Foundation of Jiangsu Higher Education Institutions of China under Grant No.08KJB520010.

References

1. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. MIT Press, Cambridge (1999)
2. Fu, X., Bultan, T.: Analysis of interacting BPEL Web service. In: Proceedings of the 13th International Conference on the World Wide, New York, USA, pp. 621–630 (2004)
3. Diaz, G., Pardo, J.J., Cambronero, M.E., Valero, V., Cuartero, F.: Automatic translation of WS-CDL choreographies to timed automata. In: Proceedings of the International Workshop on Web Service and Formal Methods, Versailles, France, pp. 230–242 (2005)
4. Zhao, X.P., Yang, H.L., Qiu, Z.Y.: Towards the formal model and verification of Web service choreography description language. In: Bravetti, M., Núñez, M., Tennenholtz, M. (eds.) WS-FM 2006. LNCS, vol. 4184, pp. 273–287. Springer, Heidelberg (2006)
5. Guermouche, N., Godart, C.: Timed model checking based approach for Web services analysis. In: Proceedings of 7th Int’l Conf. on Web Services, Los Angeles, USA, pp. 213–221 (2009)
6. Peng, Y.B., Ye, L., Zheng, Z.J.: Automatic service composition verification based on Pi-Calculus. In: Proceedings of the International Conference on E-Business and Information System Security (2009)
7. He, Y.L., Rong, M., Zhang, G.Q.: Model checking of Web service composition based on UPPAAL. Computer Science (2010)
8. Yan, Y.H., Pencole, Y.: Monitoring Web service NetWorks in a model-based approach. In: Proceedings of the 3th European Conference on Web Service, Vaxjo, Sweden (2005)

Development of LMS/LCMS (Contents Link Module) Real-Time Interactive in Videos for Maximizing the Effect of Learning*

Junghyun Kim[1], Doohong Hwang[2], Kangseok Kim[3],
Changduk Jung[4], and Wonil Kim[1,**]

[1] College of Electronics & Information Engineering at Sejong University, Seoul, Korea
junghyun64@sju.ac.kr, wikim@sejong.ac.kr
[2] Department of Information & communication at Hanyang University, Seoul, Korea
michelmk77@paran.com
[3] Department of Knowledge Information Security at Ajou University, Suwon, Korea
kangskim@ajou.ac.kr
[4] Department of Computer and Information Science at Korea University, Korea
jcd1234@korea.ac.kr

Abstract. Since most existing LMS/ LCMS have a limitation in providing various interactive elements of video contents, it lacks real-time mode and interactivity between teacher-learner and learner-learner. It is also difficult in measuring the accurate progress rate of learners in the process of teaching-learning. In this paper, we proposed a contents link technology for real-time interaction by adding various functions to videos in order to overcome limitations in e-Learning. We implemented a platform for video contents operation using the proposed method. It will overcome the existing problems and maximize the effect of e-Learning.

Keywords: Interaction, Real-time, Video-Contents, SCORM, LMS/LCMS.

1 Introduction

With the advance of computer and information communication technologies, we are living in rapidly changing information society. In the society, knowledge is being generated and terminated speedily and the role of information and knowledge is becoming increasingly important [1]. As the goals, contents and methods of education are being changed and people's interest and demand are evolving continuously, the need of knowledge is getting stronger. In response to the need, the transmission and education of knowledge are being emphasized through the spread of knowledge learning [3]. Recently, what is more, not only students but also all people are demanding new knowledge, and a solution to meet the demand is e-Learning that combines Internet (network), which is the foundation of information society, with traditional education

* This paper was conducted with the research fund for a research task (Task No. JP091004) under the Seoul R&BD Program.

** Corresponding author, +82-2-3408-3795.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 444–451, 2011.

systems. Learning in new educational environment like e-Learning can reduce the expense of education and, at the same time, provide education dynamically to anybody at any time in any place without time and space limitations. With these advantages, e-Learning is being used widely throughout the world [2, 4].

The introduction of new concepts like Web 2.0 has brought many changes in e-Learning environment. According to O'Reilly's seven principles of Web 2.0, technological trends in the age of new e-Learning 2.0 based on Web 2.0 include the expansion of SNS (Social Networking Service) and the use of blogs and podcasting in the field of education. To keep up with these trends, we need LMS(Learning Management System) / LCMS(Learning Contents Management System)-based e-Learning environment that allows learners to direct their own learning in a more natural way than existing e-Learning [5, 6]. Fig 1 shows the evolution of e-Learning paradigm.

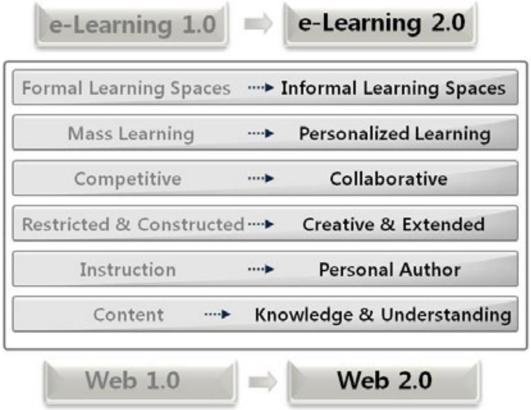


Fig. 1. Evolution of e-Learning paradigm

Therefore, this study develops a contents link technology for real-time interaction by adding various button functions to videos in order to overcome limitations in existing e-Learning, and suggests a platform for video contents operation in order to solve existing problems and to maximize the effect of learning in learners.

This paper is organized as follows. We discuss the trends and prospects of e-Learning in Chapter 2. Chapter 3 presents the proposal/design of LMS/LCMS we developed. Chapter 4 briefly describes the system implementation and a variety of functions of the system. Finally we conclude with a brief discussion.

2 Trends and Prospects of e-Learning

e-Learning technology is used widely throughout the world. In order to overcome limitations in existing learning methods and improve e-Learning systems, it is crucial to investigate existing learning technologies.

In most of e-Learning systems, LMS becomes the basic technology, and learning environment based on videos and Flash is provided. As shown in Fig. 1, e-Learning is attracting people's attention throughout the world with the introduction of the new paradigm of educational environment (e-Learning 2.0). Furthermore, it is spotlighted as a new knowledge industry and expected to produce substantial results in developing human resources. Continuous investment of money and efforts is being made in the development of e-Learning technologies and many studies are being conducted for developing new technologies for next-generation e-Learning [4, 7].

Table 1. Trends of e-Learning in 2009 and 2010

Twitter	Learn through social networking
Google WAVE **MS Sharepoint**	Build learning contents through cooperation among learners or between the teacher and learners
Rapid Learning	Modularize basic learning functions for immediate use
Mobile Learning	Enable the service of mobile products by mobile companies
Cloud Computing	A paradigm that stores information in a server permanently and maintains it in clients temporarily

Table 1 shows the trends of e-Learning in 2009 and 2010. Based on these trends, we can predict the trends after 2010 as follows. The trends include (1) 'social learning' with intensified social interaction, (2) 'smart learning' that understands the learner's characteristics and situation, the properties of the used device, etc., (3) 'exodus learning' for free learning without rules, (4) 'open learning' enhancing the value of participation, sharing and development, (5) 'synchronous learning' enjoyable all together, (6) 'just-in-time learning' that makes necessary learning available immediately, and finally (7) 'game-based learning' for absorbing natural learning [12].

These changes and prospects of paradigm suggest that existing learning environment based on videos and Flash is still insufficient in supporting interactive video contents in real-time. This raises a problem to be solved. That is, the limited environment of existing e-Learning that provides only one-way video education should be improved into learning environment that allows real-time interaction. In response to the changes of paradigm as listed above, we need to develop technologies for increasing learners'satisfaction and, at the same time, maximizing the efficiency and effect of learning.

3 System Proposal and Design

3.1 System Proposal

Existing LMS/LCMS for managing e-Learning systems has a limitation in providing various interactive video contents due to its low real-time and interactivity between teacher-learner and between learner-learner in video contents operation. Furthermore, it is hard to measure the learners' progress accurately during the learning process.

Insufficient volume or distribution of video contents may be a reason for these problems, but the above-mentioned shortages may hinder the functions of contents and the effect of learning which result in restrictions on the operation of learning programs. In order to overcome these limitations, we need to build a learning (contents) management system that supports contents in real-time and to maximize the effect of learning by developing technologies for linking contents through interactive videos.

This study designs lecture contents (LMS) according to the SCORM standards, and develops Contents Direct Link Module (LCMS) that applies various learning contents. It achieves these by promoting active communication between teacher and learner and by adding various interactive elements. Through the developed module, we expect to share video contents, to exchange information bi-directionally by chatting and replies, to enable learners to have real-time communication among them, and to keep statistics on all interactive elements occurring in the system. These functions are implemented as buttons on the video screen, and execute tasks as in the Table 2 below.

Table 2. The functions of the developed system

Real-time edition and distribution	Edit contents in real-time using the editing function, and share the contents with all users immediately
Communication & Bidirectional information transmission	Use chatting, replies, and bidirectional information exchange in real-time among users who are using the same video
Real-time statistics	View useful statistics on image view, edition, recommendation, link registration, etc.

Through the development of contents link technology, we propose a platform for video contents operation to solve existing problems. By connecting LMS and LCMS, we can support personalized learning environment and provide various real-time interactive functions including online-offline hybrid learning environment, management of learners' progress rate, and real-time learning events.

3.2 System Design

The proposed system contains functions shown in Fig. 2, which are implemented in buttons linked to LMS/LCMS on the video screen.

The system is designed to have three parts (architecture, database, and applications), and has three-tier structure of Web server, Web application server and database system, the basic conventional architecture, for processing services through the Internet efficiently.

The Web server is the main system for portal services, and functions as a gateway for accessing other systems. Business logic processing application programs for portal services provide services to Web application servers, and activate and reinforce their functions through Webtob, etc. The Web application server supports the application services of the portal system and unit systems, and enhances the utilization and performance of resources by utilizing the characteristics of JEUS and supporting JBDC.

The DB server has characteristics such as the parallel search and processing of Oracle 10g, multithreads, various backup and recovery functions, and monitoring and tuning functions. The system designed as above is built and executed as a LMS/LCMS system through which learners can learn in Web 2.0 and e-Learning 2.0 environment based on the existing functions of LMS.

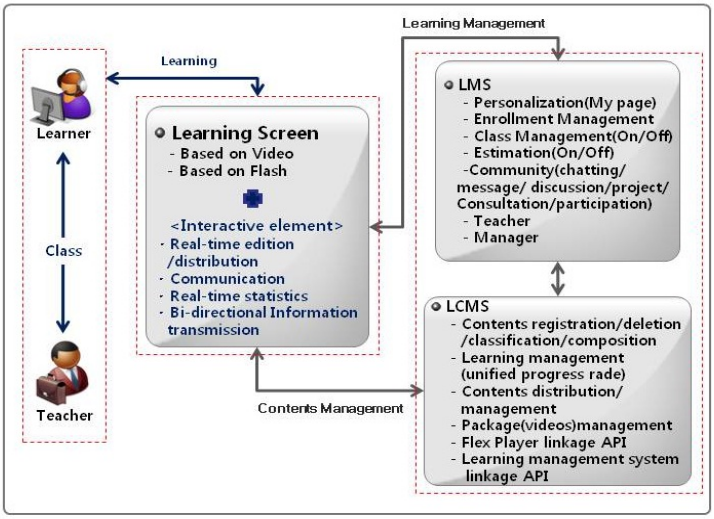

Learning Management
Learning
Learner
Class
Teacher
Learning Screen
- Based on Video
- Based on Flash
<Interactive element>
· Real-time edition /distribution
· Communication
· Real-time statistics
· Bi-directional Information transmission
Contents Management
LMS
- Personalization(My page)
- Enrollment Management
- Class Management(On/Off)
- Estimation(On/Off)
-Community(chatting/ message/ discussion/project/ Consultation/participation)
- Teacher
- Manager
LCMS
- Contents registration/deletion /classification/composition
- Learning management (unified progress rade)
- Contents distribution/ management
- Package(videos)management
- Flex Player linkage API
- Learning management system linkage API

Fig. 2. Schematic diagram of system functions

4 System Implementation and Functions

4.1 System Implementation

The system was implemented by AJAX, a Web 2.0 technology, based on aspect-oriented programming. In order to make it platform-independent, the whole architecture was designed on J2EE. Furthermore, it was designed to support environment that learners can access all services through a Web browser at any time and in any place.

Specifically, the presentation part used technologies such as HTML, JAVA Script, JSP, AJAX, DWR, and RSS/XML, and maintained the quality of sources by formulating framework for supporting integrated functions and standards. Below the area were installed common function modules based on Spring Framework, an open source supporting AOP, and core modules were developed for processing core logics and implemented in a way of guaranteeing stability and flexibility. The database part used the MS-SQL version that supports Oracle, which is the most common database software, and supports small-size operations. In the designed and implemented system, events occur in real-time through the link of a running video to the database while the user is using the system.

We implemented real-time edition and distribution functions that allow the real-time edition of contents using the contents link function (real-time learning event) and immediate sharing of the contents with other users. We created communication environment so that users can participate and share in Web 2.0 using chatting, replies, etc. in real-time instead of just viewing images, and implemented a system for viewing statistics on image view, edition, recommendation, link registration, etc. Furthermore, we built mutually cooperative real-time LMS/LCMS by implementing a system that can provide statistics on the contents progress rate and forms of learning as shown in Fig. 3.

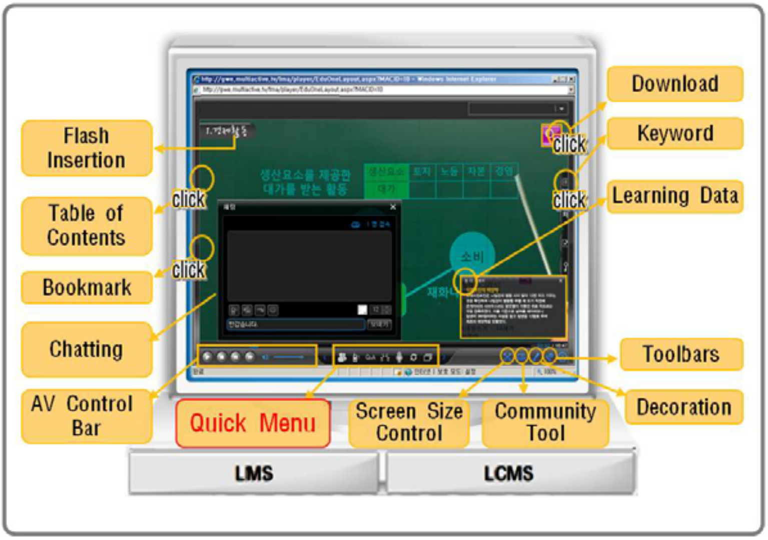


Fig. 3. Real-time video contents learning window

4.2 Various Real-time and Interactive Elements of the Implemented System

4.2.1 Real-time Edition and Distribution

Using the editing function shown in Fig. 4, the user can edit contents in real-time and share them with all the other users immediately.

This function enables the service of video learning contents based on Web 2.0 where all tasks including edition and execution can be done on the Web. Moreover, a number of teachers and learners can edit, add, delete and update contents (texts, images, Flash, sounds, videos, etc.) by working together, and distribute the outcomes of edition freely as shown in Fig. 5.

4.2.2 Communication

The communication function shown in Fig. 6 provides services such as real-time chatting, replies, bidirectional information exchange (file transmission), discussion, and Q&A, and these services promote users' participation and sharing in Web 2.0, not just viewing a video.

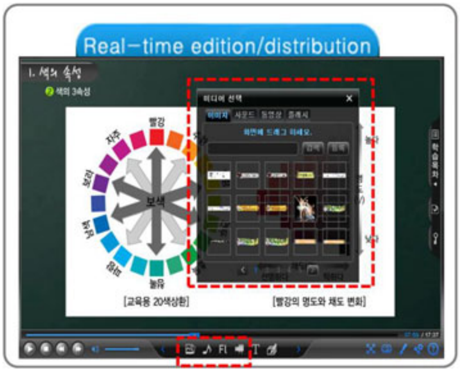


Fig. 4. Real-time edition and distribution

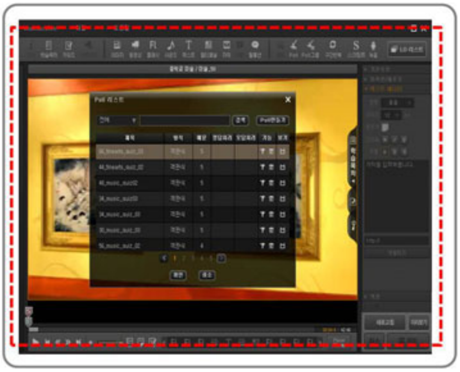


Fig. 5. Production and service functions

For one-to-one communication with an online tutor, the system provides a communication channel for conversation between the tutor and a learner who has connected at a specific time while the learner watches a video. In the communication window, functions such file transmission and walkie-talkie are available as shown in the figure above of Fig. 6-7. Furthermore, the system provides real-time functions for sending messages and related materials to multiple receivers at once as shown in the figure below of Fig. 6-7.

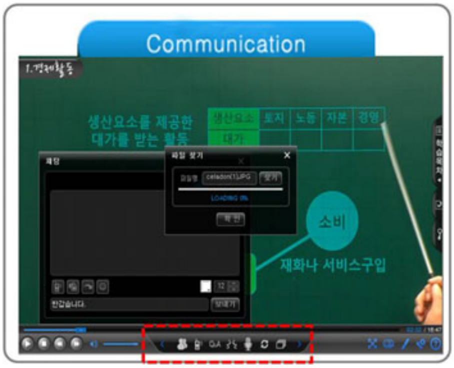


Fig. 6. Communication function(File Transfer)

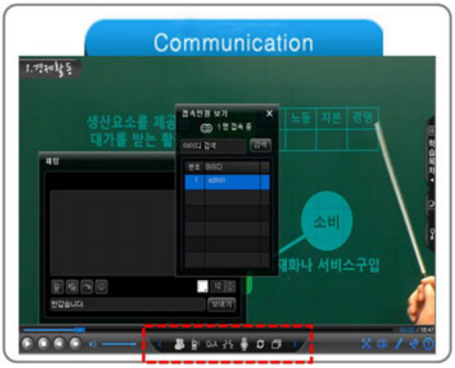


Fig. 7. Communication function(Chatting)

This function allows the users to view useful statistics on image view, edition, recommendation, link registration, etc. This function analyzes users' use of various interactive elements included in learning contents and their participation in quizzes, discussions, etc., and provides useful statistics to the users. Using these statistics, we can find and correct shortcomings of existing video contents, and manage the learners' progress rate based on their use of the system.

5 Conclusion

Changes in the e-Learning market triggered by the emergence of Web 2.0 have a very important meaning in that learners' participation begins to be spotlighted as a new factor of competition. Like the business model of Web 2.0 has exerted a huge impact on the existing market structure, e-Learning 2.0 is also increasing the importance of the construction of SNS, a participant community, in addition to the accumulation of contents. Accordingly, differentiated SNS service is expected to rise as a major factor for competition among involved companies [5, 6].

This study examined learner-centered LMS/LCMS and proposed a new contents module. The proposed method, by which learners can interact in real-time while watching a video, enhanced the effect of learning and enabled the implementation of a system for learner-oriented customized learning. Through the contents link technology, we proposed a platform for video contents operation and solved existing problems. The LCM/LCMS-based technology is expected to be the core of next-generation education for overcoming time and space limitations. It will promote the development of real-time interactive cyber learning tools.

References

1. Toffler, A.: Revolutionary Wealth: How It Will Be Created and How It Will Change Our Lives, New York (2007)
2. Anderson, C.: e-Learning: the Definition, the Practice, and the Promise, IDC (2000)
3. Drucker, P.: Webucation, Forbes (2000)
4. Li, B.: Information Technology and Its Application in e-Learning, pp. 293–296 (2009)
5. He, S., Wang, P.: Web 2.0 And Social Learning in a Digital Economy, pp. 1121–1124 (2008)
6. O'Reilly, T.: What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software (2005)
7. e-Learning consortium, e-Learning Symposium, Japan (2007)
8. Schluep, S., Ravasio, P., Schlr, S.G.: Implementing learning content management. In: Ninth IFIP TC13 International Conference on Human-Computer Interaction, Zurich, Switzerland, pp. 884–887 (2003)
9. LMS, Learning Management System, `http://www.wikipedia.org/`
10. Greenberg, L.: LMS and LCMS: What's the Difference? (2002)
11. The SCORM Overview, Sharable Content Object Reference Model (SCORM) Version 2004, ADL (2004)
12. Park, H.-J.: e-Learning expectation Trend (2010), `http://www.heybears.com/`

Converting XML Schema Data to Object-Relational Data with DOM

Lijun Sang, Jihai Xiao, and Xiaohong Cui

College of Textile Engineering and Art, Taiyuan University of Technology,
030600 JinZhong, China
slj9988@163.com, {xiaojihai,tycuixiaohong}@126.com

Abstract. There are many strategies for storing XML documents, using (object-) relational database to store XML document is one kind of the strategies. The paper discussed the new mapping method from XML Schema to the object-relational database and discussed its basic mapping rules and an approach based on DOM-Chart model in details. This mapping process preserves the structure as well as the semantic constraints of the source schema in the target schema. The test result shows that the theory and the design are feasible and effective.

Keywords: XML Schema, object- relational database, DOM-Chart, mapping approach.

1 Introduction

There are many strategies for storing XML documents, using (object-) relational database to store XML document is one kind of the strategies. So far, researchers have done a lot of work in transforming XML DTD to (object-) relational pattern, such as: the object-relational pattern algorithm[1],[4],the information-preserving algorithm[2],the path-based algorithm[3],the cost-based algorithm[5],[6],etc. But DTD has some fatal shortcomings, for example: it didn't use syntax rules of XML, it didn't support complex data type or user defined type. However, XML Schema just solves these problems.

The paper discussed the new mapping method from XML Schema to the object-relational database and discussed its basic mapping rules and the approach based on DOM-Chart model in details. The new DOM-Chart model is defined according to the modification of the DOM Tree model. The primary mapping algorithm in which the target schema is described in concept model is drawn up according to the new DOM-Chart model. The steps of converting concept model to physical model and the steps for storing XML data in object-relational database are also listed briefly. The experiment result shows that the theory and the design are feasible and effective, and that the mapping process preserves the structure as well as the semantic constraints of the source schema in the target schema.

The paper is structured as follows: Section 2 briefly sums up the related works. Section 3, 4 describes the algorithm itself in detail. Section 5 introduces its prototype implementation and presents analysis of statistics result. Finally, conclusions are provided in Section 6.

Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 452–460, 2011.

2 Related Works

There is a lot of techniques for Store XML data using (O)R databases. Generally these techniques can be classified according to several different criteria. The primary methods can be classified according to the type of documents. These methods are dividing XML documents according to their content, structure, and supposed use into data-centric and document-centric (Vakali, 2005)[7]. Another methods result from the basic ideas of the mapping methods and consists of three classes-generic, schema-driven and user-defined (Amer-Yahia 2004)[8]. Schema-driven mapping can be further divided either according to the source schema (DTD, XML Schema) or the target schema(relational or object relational) .

3 Schema-Driven Mapping

Schema-driven mapping methods are based on the existing schema S_1 of stored XML documents, written in DTD or XML Schema, which is mapped to (O)R database schema S_2. The data from XML documents in accordance with S_1 are then stored into relations of S_2.The purpose of these methods is to create optimal schema S_2, which consists of reasonable amount of relations and whose structure corresponds to the structure of S_1 as much as possible. All of these methods try to improve the basic mapping idea “to create one relation for each element composed of its attributes and to map element and sub-element relationships by using keys and foreign keys”. Schema-driven methods have some basic mapping rules:

1) Sub-elements with max occurs = 1 are mapped to tables of parent elements.
2) Elements with max occurs > 1 are mapped to separate tables. Element and sub-element relationships are mapped using keys and foreign keys.
3) Alternative sub-elements are mapped to separate tables or to one universal table (with many null fields).
4) If it is necessary to preserve the order of sibling elements, the information is mapped to a special column.
5) Elements with mixed content are usually not supported.
6) A reconstruction of an element requires joining several tables.

4 Mapping XML Schema to Object-Relational Pattern

In this part, the mapping approach is discussed in detail, The mapping rules from XML Schema to Object Relational Pattern are summarized in section 4.1; The new DOM-Chart model is defined according to the modification of the DOM Tree model, At the same time, the cycle problem in DOM-Chart is discussed in section 4.2 ; According to the new DOM-Chart model, the primary mapping algorithm in which the target pattern is described in concept model is drawn up in section 4.3; The steps of converting concept model to physical model and the steps for storing XML data in object relational database are listed briefly in section 4.4 and section 4.5.

4.1 Mapping Rules

The mapping rules from XML Schema to Object Relational pattern are as follows:

1) built-in and user-defined simple type → corresponding database simple type, together with corresponding integrity constraint(s).
2) complex type and model group → OR user-defined type (UDT), including:

XML attributes → UDT attributes with corresponding simple types,
simple element content → UDT attribute with corresponding simple type,
element and sub-element relationship → UDT attribute, whose type is (according to the allowed occurrence and the type of the sub-element) either the UDT of the sub-element or the REF/ARRAY of REF to the UDT.
3) deriving of complex types → UDT inheritance.
4) element (according to its type and allowed occurrence) → its own table or a column of the table which corresponds to its parent element.
5) root element → an element having a complex type without attribute.
6) the constraints of complex type are not discussed.
7) the sequence of items→UDT seq, whose type is determined according to the type of the item. The element-sequence relationship depends on the maximum occurrence and the type of the whole sequence.

If the element occurs many times, the order of sibling elements is maintained by arrays or indexes; if the element occurs in a set, it's order is preserved by auxiliary attributes; the mixed content is preserved by storing the mixed-content elements as other elements or their text parts in a separate multi-valued property.

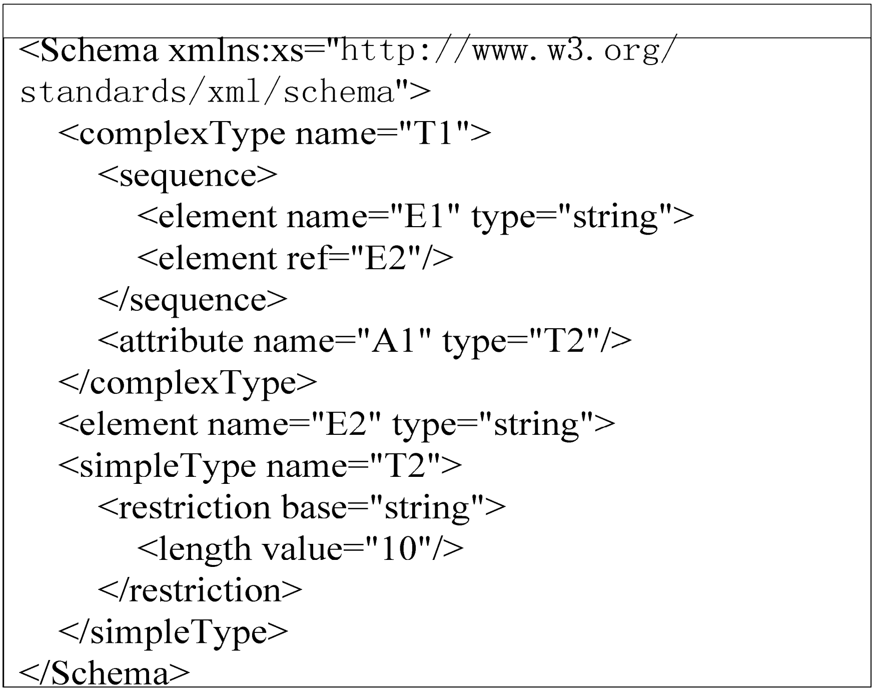
4.2 DOM-Chart Model

A new model is defined according to modification of the DOM Tree model, and it is named DOM-Chart in brief. Next, on the basis of the DOM-Chart and the mapping rules, a new algorithm is put forward in section 4. This section includes the definition of the DOM-Chart and exploring a solution to the cycle problem in DOM-Chart.

Let $T_{DOM}=(V_T,E_T)$ which is the DOM Tree model of the given XML-schema document, and $Glob^T{}_{type} \subseteq V_T$ is the subset of the global defined simple or complex type, $Glob^T{}_{item} \subseteq V_T$ is the global defined element, element group, attribute, attribute group. So the DOM-Chart is defined as follows:

Definition 1. $G_{DOM}=(V_G,E_G)$ Here:
$V_G=V_T$
$E_G=\{ (V_x,V_y) | \{V_x,V_y\} \in E_T \wedge$ (element node V_y is a child of the element node V_x) $\} \cup \{ (V_x,V_y) | \{V_x,V_y\} \in E_T \wedge$ (attribute node V_y is a attribute of the element node V_x) $\} \cup \{ (V_x,V_y) | V_x$ is a attribute node whose type refers a simple or complex type V_y node, $V_y \in Glob^T{}_{type} \} \cup \{ (V_x,V_y) | V_x$ is a 'ref' node referring V_y node, $V_y \in Glob^T{}_{item} \}$

Thus, the DOM-Chart is composed of all node in DOM Tree and two kinds of edges including the original edges and the auxiliary edges denoted by solid lines and dash lines respectively in figure 2. In addition, the letter **E** in brace denotes element, and the letter **A** denotes attribute.



```
<Schema xmlns:xs="http://www. w3. org/
standards/xml/schema">
  <complexType name="T1">
    <sequence>
      <element name="E1" type="string">
      <element ref="E2"/>
    </sequence>
    <attribute name="A1" type="T2"/>
  </complexType>
  <element name="E2" type="string">
  <simpleType name="T2">
    <restriction base="string">
      <length value="10"/>
    </restriction>
  </simpleType>
</Schema>
```

Fig. 1. An example of XML-Schema document

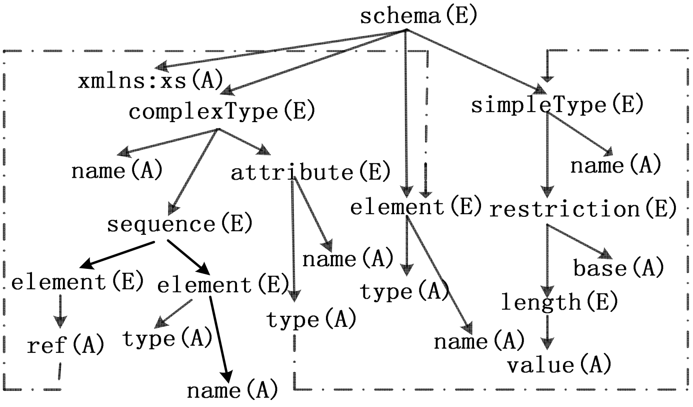


Fig. 2. The DOM-Chart model corresponding to figure 1

4.2.1 Solution to the Cycle Problem in DOM-Chart

The cycle in DOM-Chart is caused by the reference among the global defined elements or the reference among complex types. In SQL:2003, it's possible to extend the type system, and most current DBMS have supported the cycle definition in UDT, such as ORACLE 10g release 2. User can make use of the incomplete type in above DBMS. Thus, after defining the incomplete type, user can use it instead of the complete type.

This problem can be solved via using the incomplete type. Before processing the DOM-Chart, an incomplete type is created for each globally defined complex type/element. Then the incomplete types are used instead of the corresponding complete types and their definitions are completed.

4.3 Mapping Algorithm

The whole mapping process is simplified by using the DOM-Chart mentioned in section 4.2. The algorithm includes three parts: 1)constructing DOM-Chart model. 2) solving cycle problem etc.3)giving recursive sub-procedure which deals with the node in DOM-Chart. The algorithm is as follows:

```
1 ) Constructing DOM-Chart model: $G_{DOM}$= ( $V_G$,$E_G$ )
Construct a DOM tree $T_{DOM} = (V_T, E_T)$ according to given XML Schema document;
For each node $v_T \in V_T$ do
   create the corresponding node $v_G \in V_G$;
For each (unordered) edge $e_T \in E_T$ do
   create the corresponding (ordered) edge $e_G \in E_G$;
Let $Glob^T_{type} \subseteq V_T$ and $Glob^T_{item} \subseteq V_T$ be defined like in Definition 1;
Let $Glob^G_{type} = \{v_G \mid v_G \in V_G \wedge v_G$ corresponds to $v_T \in Glob^T_{type}\}$ and
Let $Glob^G_{item} = \{v_G \mid v_G \in V_G \wedge v_G$ corresponds to $v_T \in Glob^T_{item}\}$;
For each base, type, itemType or ref attribute node $a \in V_G$ referencing to node $g \in$
$Glob^G_{type} \cup Glob^G_{item}$ do
   create the corresponding (ordered) $e = (a, g) \in E_G$;
2 ) Generating the UDT and table etc.
Mark each $v \in V_G$ as unprocessed;
Let $Glob^G_{elem} \in V_G$ be the set of globally defined elements;
For each $v \in Glob^G_{type} \cup Glob^G_{elem}$ do begin
   Create its incomplete UDT;
   Mark v as processed;
End
ProcessNode(Schema root node $\in V_G$);
For each $v \in Glob^G_{type} \cup Glob^G_{elem}$ with an incomplete UDT do
   create its complete UDT;
For each element, which is not mapped to a typed column do
   create a typed table (including corresponding column integrity constraints);
For each reference (except for those corresponding to a choice of elements) or array
of references do
```

create a corresponding SCOPE integrity constraint;

3) Recursive sub-procedure

```
procedure ProcessNode($v \in V_G$)
begin
   Let $V_{child}$ be the set of child nodes of v;
   For each $v_{ch} \in V_{child}$ do begin
      If $v_{ch}$ is marked unprocessed then ProcessNode($v_{ch}$);
      Else if ($v$, $v_{ch}$) $\in E_G$ expresses an extension of $v_{ch} \in Glob^G{}_{type}$ and $v_{ch}$ has an incomplete UDT then begin
         ProcessNode($v_{ch}$);
         Create its complete UDT;
      end
      Else if v is a Schema root node and $v_{ch}$ has an incomplete UDT then ProcessNode($v_{ch}$);
   end
   If v is marked unprocessed then
      process v according to the established mapping rules;
   Mark v as processed;
End;
```

According to the above algorithm, all the items, including UDTS, type tables, references, constraints, are considered in the target pattern. Moreover, the cycle problem is solved by using the incomplete type.

4.4 Convert Concept Pattern to Physical Pattern

The algorithm of mapping from XML Schema to object-relational creates object-relational pattern irrelevant to RDBMS. Here, it is called concept pattern. In theory, after concept pattern is created, the conversion from XML Schema to target pattern is finished. In order to complete the experiment and to research further, this section finished the conversion from concept pattern to physical pattern. The steps of producing physical pattern are as follows:

Step 1: According to the depiction of concept pattern, by building the mapping table between data type in concept model and DBMS data type, and by finding the default values and constraints of field in concept model, then the DDL sentences are synthesized.
Step 2: In order to ensure entity integrality, primary key is created, and the DDL sentences are synthesized.
Step 3: According to the relationship between the element and sub-element, the foreign key is created to ensure reference integrality, and to synthesize DDL sentence.

4.5 Store the XML document to ORDB

The steps of storing an XML document to object- relational database are as follows:
Step 1: To parse an XML document and create DOM Tree model on the basis of the DOM model.

Step 2: To ascertain fields in object-relational pattern for the data in the XML document.
Step 3: To extract the leaf node of DOM Tree and to supplement the data of auxiliary fields.
Step 4: To synthesize the SQL sentences and to insert data into tables.

The pattern structure of XML document is considered in the whole storage process, at the same time, auxiliary information in OR pattern is also thoroughly considered. Auxiliary information is stored into another table in the mapping process.

5 Experiment

The experiment uses DBMS Oracle 10g database which supports SQL: 2003 Standard and object-oriented characteristics. It uses JAVA to implement the concrete algorithm. Because XML Schema document that is described by XML Schema language is still an XML document, it is parsed by the way that is used to parse XML document such as using the JDOM program interface.

5.1 Experiment Route and Data

Experiment route is depicted in figure 3, and the experiment implements two parts as follows:

(1) To create object-relational schema according to the given XML Schema
(2) To store XML document data into object-relational tables.

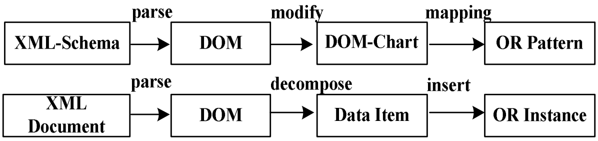


Fig. 3. The experiment route

Experiment data includes three data sources:

(1) W3CXML: it is a standard example used to depict XML-Schema for W3C.
(2) IBMXML: it is an XML-Schema document that describes the resume format in IBM Corp.
(3) XMLSpy1, XMLSpy2, XMLSpy3: these schemata and data such as OrgChar.xsd, DataSheet.xsd, Address.xsd can be found in software XMLSpy2008.

5.2 Experiment Result and Statistics

Statistics of experiment are depicted in table 1.

Sour(Sources) denotes Data Source.
Rela(Relations) denotes the number of the relations in object-relational pattern.

Attr(Attributes) denotes the quantity of attributes in object-relational pattern.
A pair of numbers in $T_i(0<i<7)$ columns denote the number of traits in source schema and the number of traits preserved in target schema respectively.
T_1 denotes data type ; T_2 denotes field value restriction; T_3 denotes default value;
T_4 denotes occurrence times; T_5 denotes unique restriction
T_6 denotes primary key and foreign key

Statistics in table1 shows that the quantity of relations in target schema is less than the quantity of elements in XML Schema, and that the quantity of attributes in target schema is larger than that of properties in XML Schema. Moreover, statistics in table1 show that a majority of semantic characteristics are preserved in target schema.

Table 1. The experiment statistics of converting XML Schema to Object-Relational Pattern

Sour	Rela	Attr	T1	T2	T3	T4	T5	T6
W3C XML	4	43	15/15	2/2	1/1	5/0	0	3/3
IBMXML	12	165	97/95	66/63	0/0	66/0	0	36/26
XML Spy1	7	85	27/27	15/11	0/0	10/0	0	0/0
XML Spy2	11	82	40/37	12/10	0/0	18/0	0	0/0
XML Spy3	2	15	7/7	2/2	1/1	2/0	0	0/0

6 Conclusion

The paper discussed a new mapping method from XML Schema to object-relational database, and discussed its basic mapping rules and the approach based on DOM-Chart model in details. The experiment result shows that the theory and the design are feasible and effective, and that the mapping process preserves the structure as well as semantic constraints of the source schema in the target schema. Comparing this new mapping method with other methods [9], it uses the XML Schema and object-relational database characteristics.

References

1. Runapongsa, K., Patel, J.M.: Storing and querying XML data in object-relational DBMSs. In: Chaudhri, A.B., Unland, R., Djeraba, C., Lindner, W. (eds.) EDBT 2002. LNCS, vol. 2490, pp. 266–285. Springer, Heidelberg (2002)
2. Barbosa, D., Freire, J., Endelzon, A.: Designing Information-Preserving Mapping Schemes for XML. In: Proceedings of the 31st VLDB Conference, Trondheim, Norway (2005)
3. Yoshikawa, M., Amagasa, T., Shimura, T., Uemura, S.: a path-based approach to storage and retrieval of XML documents using relational databases. ACM Transactions on Internet Technology, 110–141 (2001)
4. Mlynkova, I., Pokorny, J.: XML in the World of (Object-) Relational Database Systems. Charles University, Prague, Czech Republic Technical report (2003)
5. Bohannon, P., Freire, J., Roy, P., Siméon, J.: From XML Schema to relations: A cost-based approach to XML storage. In: Proceedings of the International Conference on Data Engineering (ICDE), pp. 64–75 (2002)

6. Hu, T.-L., Chen, G.: Adaptive XML to relational mapping: an integrated approach. Journal of Zhejiang University Science A, 6–9 (2008)
7. Vakali, A., Catania, B., Maddalena, A.: XML data stores: emerging practices. Internet Computing 9(2), 62–69 (2005)
8. Amer-Yahia: Overview of Existing XML Storage Techniques. AT&T Labs, Cambridge, UK. W3C Recommendation, `http://www.w3.org/TR/xmlschema-2/`
9. Sihem, A.Y., Fang, D., Juliana, F.: A Comprehensive Solution to the XML to Relational Mapping Problem. In: WIDM 2004, Washington, DC, USA (2004)

XML Query Algorithm Based on Matching Pretreatment Optimization

Yi Wang[1], Heming Ye[1], Haixia Ma[1], and Weizhao Zhang[2]

[1] Higher Education Research Institute of Training Department, Ordnance Engineering College,
No.97 Hepingxilu Road,
050003 Shi Jiazhuang, Hebei, China
[2] Administration Office of Training Department, Ordnance Engineering College, No. 97
Hepingxilu Road,
050003 Shi Jiazhuang, Hebei, China

Abstract. In the XML query algorithm based on matching pretreatment, three important existing tree matching models are used and the match results of data sets are presented according to the match cost in a descending order. The function of matching pretreatment is added to improve the existing algorithm, and a series of experiments are conducted. The results show that this algorithm can remove the unwanted node in the tree to promote the efficiency of data sets retrieval especially the recall ratio, the precision ratio and the average response time, when the data scale is quite large. This algorithm is applied in the unity retrieval system of scientific and technological resources database. It can facilitate the navigation of resources, narrow the search scope and promote the efficiency of this system.

Keywords: XML; Tree matcher Models; Matching pretreatment.

1 Introduction

XML is rapidly developing to be the standard for the data representation, integration and exchange in the Web. Lots of Metadata based on XML have been used at present. However, Methods of metadata retrieval for the resource subjects use now as follows. Either the traditional database retrieval technology or the traditional retrieval based on the key words matching for the library and the cataloguing query technology for the information science [1] is used. Above the methods only can search the results which are exactly matching with the query conditions, thus the recall ratio of the resource subjects is very low.

In the recent years, scholars had proposed a lot of retrieval technologies and proposals based on the XML matching. Richterl[4] proposed a query method of XML files based on structure. Although this method can get high recall ratio, it is applied for the ordered trees. This method is not suitable for the metadata query on XML, because the trees on XML are the unordered trees. Mr Xu[5] proposed a software components query and matching algorithm based on the inclusion matcher for unordered trees. The algorithm can remarkably promote the recall ratio for the components, and provide the boolean query service. But this algorithm didn’t work

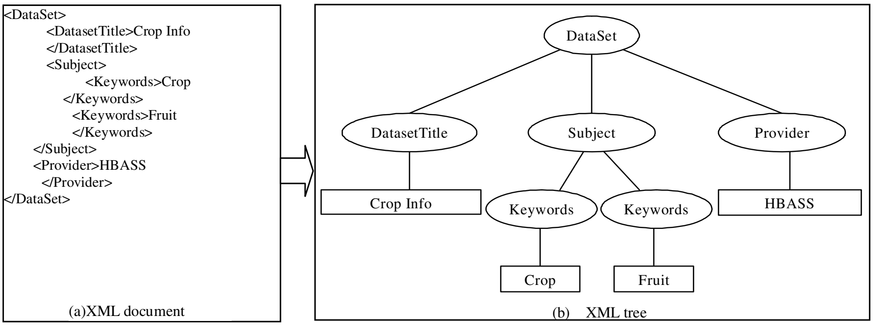
Z. Gong et al. (Eds.): WISM 2011, Part II, LNCS 6988, pp. 461–467, 2011.

well when many unused nodes are existing. Aiming at this situation, this paper proposes a metadata query algorithm based on pretreatment, which introduces the tree matching theory into the XML metadata query and applied in the unity retrieval system of technology and science databases in Hebei province. It is proved that this algorithm can serve well for the resource navigation in the system, narrow the search scope and make the system use easily.

2 Basic Conception and Matcher Models

2.1 XML Trees

XML trees are the trees in which u and v are the two random elements, u' and v' are the two corresponding nodes in the XML tree, u is equal to v when u' is equal to v', and u includes v when u' is the ancestor of v'.



```
<DataSet>
        <DatasetTitle>Crop Info
        </DatasetTitle>
        <Subject>
               <Keywords>Crop
           </Keywords>
             <Keywords>Fruit
             </Keywords>
      </Subject>
      <Provider>HBASS
        </Provider>
</DataSet>
```

(a)XML document

(b) XML tree

Fig. 1. This shows a XML document and its XML tree.Figure 1(a) is an example of XML document, and Figure 1(b) is an XML document structure tree of it. The ellipse node in the XML tree presents an element in the XML document, and rectangle frame presents the values of it.

2.2 Tree Matcher Models

Embedding Tree Matcher. Embedding tree matching [10] demands that the querying nodes keep the relationship of parent and child, and the extra nodes are not allowed to exist in the same time. Therefore, the Embedding matching can be regarded as the exact matching. The precision ratio is high when using embedding tree matching, but the recall ratio is very low. The embedding tree matching models are shown in the figure 2(a).

Tree Inclusion Matcher. Tree inclusion matcher relaxes the corresponding relationship among nodes. It only requires that the nodes pairs keep the relationship among ancestor and posterity, and the extra nodes are permitted when matching the inclusion tree. So the recall ratio is promoted when matching the inclusion tree. This is shown in the figure 2(b).

Tree Loose Inclusion Matcher. Tree loose inclusion matcher relaxes the corresponding relationship among nodes again. It not only permits the existence of useless nodes in the queried tree, and also allows that the tree to be queried can be lack of useful nodes. The matching nodes just have the approximate relation. The improvement enhances the ability of fuzzy query based on exact matching, and increases the recall ratio. This is shown in the figure 2(c).

The process of from embedding tree to inclusion tree then to loose inclusion tree is that the recall ratio is gradually increasing and the precision ratio is partially decreasing. These three matcher models, each has own practical place. They can cooperate and coordinate mutually, and provide the individual query service for XML documents which can be fine-tuning jointly.

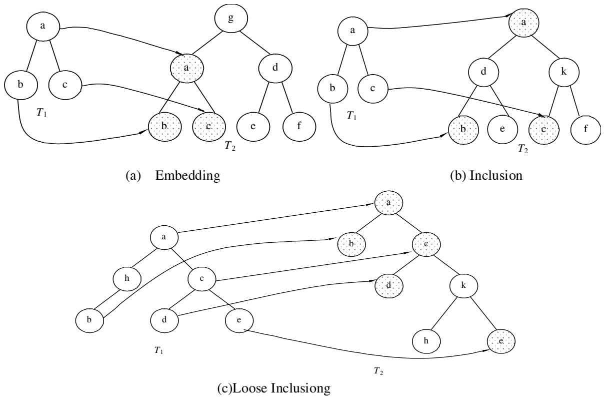


Fig. 2. This shows the three tree matching models. Fig. 2(a) is an embedding tree, Fig. 2(b) is an inclusion tree, Fig. 2(c) is a loose inclusion tree.

3 Xml Query Based on Matching Pretreatment

3.1 Matching Pretreatment

The data sets objects are described by the same metadata standard which is the core metadata standard of science resources. Every node of the metadata tree of data sets objects is certain part of metadata tree. The differences are the elements values of leaves nodes.

The internal nodes in the data sets objects such as D_1, D_2 and D_3 come from the metadata tree A. The process of pretreatment is shown in figure 3. The content of leaves nodes in the information object metadata tree which is the element value is depict in detail. Hence, Doing the matching operation between the query tree and metadata tree before query, and pretreatment information in the process is recorded. The pretreatment information is used to save the time spent in the process of unrelated nodes matching operation between query tree and metadata tree and the repetitive conditions about the unrelated nodes. The detailed algorithm is shown in figure 4.

For example, the node b_1 in the query tree is matching the nodes of a_{11} or a_2 in the tree A, and the node b_2 in the query tree is matching the node of a_{12} in the tree A. The information above is the pretreatment information. It is unnecessary to do the matching operation between query tree Q and metadata tree of object D_2 because of the inexistence of nodes which can match the nodes in tree Q. Moreover, it is unnecessary to consider about the matching condition between Q and sub tree whose root is node a_3 because this sub tree doesn't have the matching nodes with Q. The efficiency of retrieval process can be enhanced through the pretreatment process introduced above.

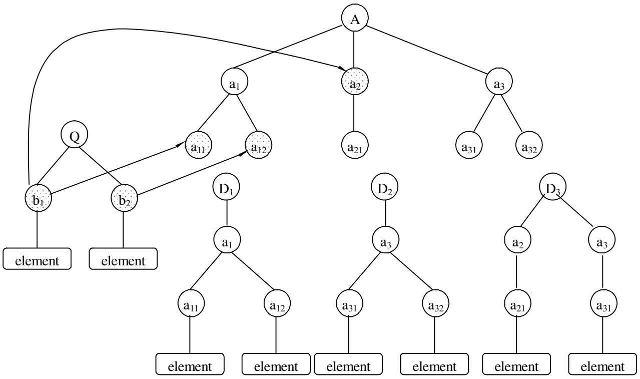


Fig. 3. This shows the process of pretreatment

```
Ⅰ is used for Pretreatment Information Acquisition.
Input ： Query Tree Q，Metadata Tree A
Output：Pretreatment Info
While(Q.subNode!=leaf)
{
  If(Find(Q.subNode,A)
    {
        Record(Info,subNode);
     }
    Else
    {
        DoNothing;
    }
  subNode.next();
}
```

```
Ⅱ is used for the omission of useless nodes
Input ：   Info ，  Destination  object  Set
D={$D_1$,$D_2$,……,$D_n$}
Output：Result object Set D={$D_1$,$D_2$,……,$D_m$}(m≤n)
For(i=1;i<=n;i++)
{
While($D_i$!=NULL)
{
If(Inclusion($D_i$.subNode,Info) {
        j[i]++; }
      Else{
        DoNothing; }
subNode.next();}}
For(i=1;i<=n;i++){
  If (j[i]==0) {
      Delete($D_i$,D); }
Else{
 DoNothing; }
}
```

Fig. 4. This shows the pretreatment algorithm

3.2 Matching Pretreatment

Matching Process is shown below:

(1) Metadata tree is generated.

(2) Query condition is analyzed, and then query tree is generated.

(3) The useless nodes are omitted after the process of matching operation between the query tree and the metadata tree. The method is shown in part 3.1.

(4) The query tree and metadata tree are encoded. The encoding principle is as follows. The node of i in layer n is encoded $a_1.a_2.a_3.a_4.\ldots\ldots a_{n-1}.i$.

(5) The regions are opened in memory. The codes of same elements nodes are putted into sets region. The results and intermediate results in the query process are putted into result region. The last query results users needing are putted into results cache.

(6) The elements nodes of metadata tree are putted into sets region.

(7) The root in the query tree is readied, then the corresponding element is looked up in the sets region and the corresponding code is readied. These information is putted into results region. The root is set as the present node.

(8) The nodes in the under layer of present node of query tree are readied. If the nodes are existed, then go to the ninth step. Otherwise, go to the eleventh step.

(9) Three steps are as follows.

①The corresponding nodes elements are looked up in the sets region. One of them is taken to match nodes in the results. If the matching principles which are introduced from the three tree matching models are satisfied, the tree and node are isolated from the results region, they are put into the results cache, and the copied original tree is put back into the results region.

②If they are not satisfying the matching principles, or the parent nodes are not existed, then the skipped generation matching operation is carried out.

Ⅰ.If they are satisfying the structural relationship in the query tree, the nodes are separated out.

Ⅱ.Doing the matching operation with other trees in the results region until the root is not satisfying the principle.

Ⅲ.If all above are not satisfying the principle, this element will be regarded as the root and separated out into the results cach.

③ These elements having the same code are collected one by one. At last the trees in the results cache are taken into the results region, in the mean while the copied trees are deleted.

(10) The brother nodes in the under layer of the present node in the query tree are readied. If they consists, then go to the ninth step. Otherwise, the node of the under layer will be set as present node, and go to eighth step.

(11) The trees which have the highest matching degree will be selected and return to users according to the matching degree of the trees in the results region calculated by the formula of matching degree.

3.3 Calculation of Matching Degree

The formula of matching degree [11]is shown in (1) towards the query tree T_0 and the result tree searched from metadata tree T_0.

$$Tmd[T_0,T_1]=\frac{|V_1|}{|V_0|}\cdot\sum a_{uv}\frac{Min(l_{uv},L_{uv})}{Max(l_{uv},L_{uv})} \quad . \tag{1}$$

Formula (1) is explained as follows. Nodes u and v are the valid nodes which should keep the relationship of ancestor and posterity in tree T_1.The parameter of l_{uv} presents the number of edges contained in the route between u and v of tree T_1. The parameter of L_{uv} presents the number of edges contained in the route between u and v of tree T_0. The parameter of a_{uv} presents the route weight between the node u and v in the tree T_0.

The quality of query results should be considered from the two aspects. One is the number of valid nodes separated from the tree. That is how much useful information is searched. The other is the structural relationship between the valid nodes. That is the depth and branches of the tree structure of query result. The formula is defined to marcato embody the two points. $\frac{|V_1|}{|V_0|}$ is for the first consideration,and $\sum a_{uv} \frac{Min(l_{uv}, L_{uv})}{Max(l_{uv}, L_{uv})}$ is for the second consideration.

4 Experiments and the Analysis of Results

Metadata query based on XML is realized by JDOM and JSP technologies in the B/S mode. The XML description is depicted by core metadata standard for the all kinds of data sets in the science and technology database of Hebei province. It is stored in XML documents, and the query requires are raised according to the definition items in the metadata standard.

Three parameters are set as follows. Parameter s presents the data scale. Parameter d presents the depth of query tree. Parameter n presents the number of nodes of query tree. Let s equals 500 and 2000. First set is that d is set as 2 and n is set as 4. Second set is that d is set as 3 and n is set as 8. Third set is that d is set as 4 and n is set as 16. The experimental results only show the statistics data of loose inclusion matching operation. They are shown in the Table1 and Table2.

The average response time is only about 300 ms when the number of query nodes is 16 and the data scale is 2000 even for the more complex loose inclusion matching operation. It is indicated that the tree matching algorithm based on matching degree in which the pretreatment is introduces is satisfying in the aspects of the recall ratio, the rate of accurate survey and the average response time.

Table 1. Performance situation when s=500

Set No.	d	n	recall ratio/%	precision ratio /%	average response time /ms
1	2	4	100	100	87
2	3	8	99.61	99.35	110
3	4	16	98.69	99.23	132

Table 2. Performance situation when s=2000

Set No.	d	n	recall ratio/%	precision ratio /%	average response time /ms
1	2	4	99.76	99.28	218
2	3	8	99.43	97.15	253
3	4	16	98.33	95.03	296

5 Conclusions

The XML query algorithm is making use of the existing three tree matching models, and the matching results for data sets is obtained on the basis of matching degree. Moreover, the algorithm is improved with pretreatment. A series of experiments are conducted. It is that this algorithm can remove the unwanted node in the tree, and promote the efficiency of data sets retrieval when the data scale is very large.

References

1. Frakes, W.B., Pole, T.P.: An empirical study of representation methods for reusable software components. IEEE Transactions on Software Engineering 120, 617–630 (1994)
2. Torshen, S.: ApproXQL: Design and implementation of an approximate pattern matching language for XML. Technical report. Free University, Berlin (2001)
3. XML and query language: Experiences and examples, `http://www-db.research.bell-labs.com/user/simeon/xquery.ps`
4. Richter, T.: A new measure of the distance between ordered trees and its applications, Technical report, University of Bonn (1997)
5. Xu, W., Qian, L., Cheng, J.: Research on Matching Algorithm for XML-Based Software Component Query. Journal of Software 14, 1195–1202 (2003) (in Chinese)
6. Torsten, S., Naumann, F.: Approximate tree embedding for querying XML data. In: Proceedings of ACM SIGIR Workshop on XML and Information Retrieval, Athens, Greece, pp. 181–184 (2000)
7. Sun, J.: Application of GA in XML Data Query. Computer Engineering 12, 183–184 (2005) (in Chinese)
8. Lu, Y., Zhang, B., Guo, J., et al.: A XML Query Solution with User-friendly Interface. Mini-Micro System 24, 1849–1852 (2003) (in Chinese)
9. Xiao, H., Tang, C., Zhang, T., et al.: BTCS: The Binary Traveling Coding Scheme for XML Document. Si Chuang University Journals (Nature Science version) 43, 532–537 (2006) (in Chinese)
10. Li, Q., Moon, B.: Indexing and querying XML data for regular path expressions. In: Proceedings of the 27th International Conference on Very Large Data Bases, Rome, Italy, pp. 1811–1814 (2001)
11. Yin, P., Wei, J., Zheng, W.: A XML-Based Approach for Information Search. Computer Engineering and Science 29, 145–148 (2007)

Author Index