# Research on Social Network Based on Meta-Search Engine

SHEN Yang
School of Information Management
Wuhan University
Wuhan, China
e-mail: yshen@whu.edu.cn

LIU Zi-tao
International School of Software
Wuhan University
Wuhan, China
e-mail:hnly228078@gmail.com

LUO Cheng
School of Computer
Wuhan University
Wuhan, China
e-mail:dolrill@qq.com

LI Ye
School of Electronic Information
Wuhan University
Wuhan, China
e-mail: 411131013@qq.com

***Abstract*—In order to solve the problem that we can only collect data from one single data source at some fixed time after mining the keywords in a rather superficial level, and to take full use of the information returned by search engines to construct the social relationship network based on the semantic link of the searched subject, we do the regular research by using the ROST Content Mining System which helps to undergo the process of page monitoring, content analysis and social network mining based on the pages returned from the four search engines (Google, Baidu, Sougou and Youdao). In the mining process, we adopt the cross-page framework adaptive algorithm which helps to solve the instability problem of the HTML framework codes, to extract information from the acquired web pages. Then we extract the cooccurrence set of high-frequency characteristic words to create the tridimensional social network graph by adopting the progressive search algorithm in the meta-search engine to extend the attribute set of the keywords. Finally, we conducted three typical case studies. They are the comparison of the coverage rate between Google and the meta-search engine, the dynamic changes in real-time network based on the meta-search engine and the progressive mining of effective content in meta-search engine, which all showed the advantages of the method in which we proposed the meta-search engine, as we could have more data sources, stronger real-time dynamic monitoring capacity, and deeper progressive searching ability. So we propose this meta-search engine method which can be used in social network study, aiming to develop the quality of the social network based on content mining, observe the hiding relationships in deeper levels and widen the research scope of content mining.**

***Keywords-social network; content mining; meta-search engine; adaptive algorithm; progressive algorithm; real-time network***

## I. INTRODUCTION

In Web 2.0 age, the social network integrates a mass of information from which we can mine information to construct one person's social network [1]. The data collected in the process of traditional information mining, comes from one single search engine, or some closed virtual communities like FaceBook. With the help of these data, we can reproduce one person's social network in the real life [2]. While, the research method based on meta-search engine which is used to study social network evolution in social network mines information on the basis of several search engines' dataset. Compared with the method based on traditional search engines, the research method based on meta-search engine has the following advantages: firstly, it has more channels to get data, covering more pages of the Internet [3] ; secondly, it can mine out the social network of figures at all times and in all over the world by searching data from different sources, which surely breaks through the limitation of only discovering the social relationships of the present persons in virtual communities [4]; lastly, the method can do progressive mining to find out deeper social and semantic relationships [5], to construct the social network based on the semantic link.

## II. RELATED WORK

Researches of progressive analysis and content mining on the data generated in the cloud computing method from huge search engines have become the hot spot in information science. Jin, Yingzi, and some other people extracted the relationships of companies through search engines and text processing tools, and then they gave suggestions on choosing business partners based on the structural advantages in social network [6,7]. Using the data set of web sites about trademarks and stars provided by the search engines, Gloor, Peter A. put forward a new method to calculate the popular degrees of trademarks and celebrities with the help of social network analysis tools [8]. Geleijnse, Gijs extracted life stories of historical figures from the multiple and structureless information sources in Web, and then obtained their internal social relationships through their co-occurrence in Web [9]. Lin, Ming-Shun and some others used Google to construct the social network of entities, so that the category of each entity and their relationships could be marked out [10]. Matsuo, Yutaka and their mates used co-occurred information searched from Google and Web documents to extract people's relationships, and then they further classified their social relationships to get the person-to-word relationship set after clustering the people and finding out each person's keyword with their self-invented social network tools' help [11]. However, these researches have some common limitations: the first is singleness of data sources, namely, these researchers only used Google to collect data; Secondly, these researchers only collected data at some point, which obviously missed out the data's dynamic evolution in social network, ignoring the difference caused by temporal variation [12]; thirdly, the researchers did no progressive searches, merely extracting and mining data from the one-time search in search engines [13]. Based on the analysis above, we decide to gather data sets from four search engines, and then monitor them

978-0-7695-3874-7/09 $26.00 © 2009 IEEE
DOI 10.1109/WISA.2009.21

every two weeks to do our differential comparison analysis, which will help us get the dynamical data sets to do our progressive searches and promote relevant researches on social network.

## III. SOCIAL NETWORK EXTRACTION

### A. Information Extraction Model

At present, we mainly use three methods to acquire information of search engines from Internet. One is to search the partly open data set of search engines; another is to call the web service provided by search engines; the other is to monitor search engines regularly to do further analysis on the returned content. Researchers can get a huge and complex data set using the first method which, however, has bad timeliness. The second method has the advantages of easy programming and accurate results and the disadvantages of not full-scale data interface services provided by all search engines, as the result of which we cannot combine the other methods' strong points. Based on the analysis above, we decide to adopt the third method. With the help of our self-made software ROST Content Mining System [14] , we monitor pages of Google，Baidu，Sougou, and Youdao regularly. After that we analyze the content and mine the social network based on the returned pages. The software's construction shows in Fig.1.

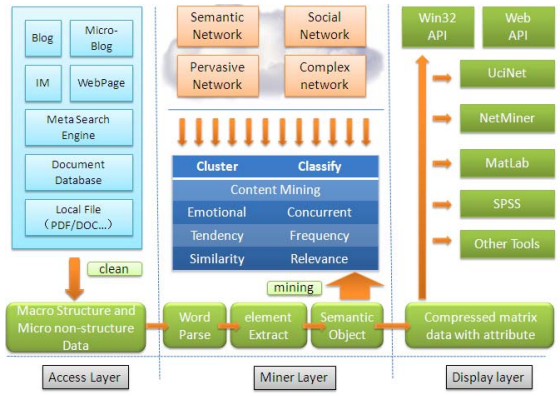


Figure 1. ROST Content Mining System Architecture

Based on the ROST CM software, we come up with a new model for social network extraction in Web. Through inputting some keyword, the software automatically analyzes the results returned by the meta-search engine, and extracts metadata such as titles, time, links, and abstracts in corresponding pages. The software will filter out the repeated pages through the web page disambiguation algorithm and then extract the abstract part in the web page to conduct the characters of Chinese language segmentation. Users can also self-define the words that need to be analyzed and mined through the help document, so that the software calculates out the relationship frequency through the co-occurrence analysis, and gets the graph of the relationship network based on the keyword. At the same time, the software can extract the phrases whose frequency is higher than some threshold and use them as seeds of the characteristic word set, which can be used to get the complete characteristic word set based on the keyword through the "progressive search algorithm". Then the high covering and low repeating social network can be constructed. The concrete model shows in Fig.2.

### B. Basic Algorithm

#### 1) Cross-Page Framework Adaptive Algorithm

Considering the fact that search engines usually return more than one web page based on one record and the HTML framework codes of the pages are usually unstable, we

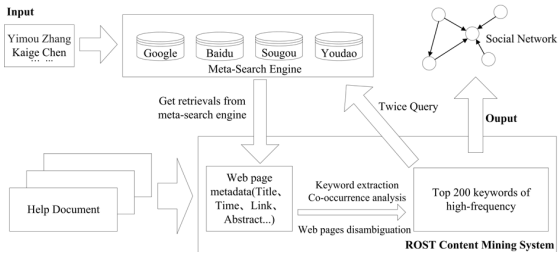


Figure 2. Extraction Model of Social Network Based on Web

propose the "cross-page framework adaptive algorithm" to achieve data acquisition in search engines.

Firstly, we define a string list as the keyword set which will be searched in search engine, and then we use ROST Content Mining System to start the search engine at certain time to query one record from the keyword set. As the returned pages based on one record are usually more than one, we do the query in sequence based on the returned pages, and store the returned pages in the web page data set in the required format. At the same time, because of the instability of the HTML framework codes, we adopt the "cross-page framework adaptive algorithm" to avoid modifying the software regularly instead of the traditional text string location method which is relatively rigid.

The cross-page framework adaptive algorithm can judge the framework of the source codes of the returned pages. Firstly the algorithm acquires the HTML source codes of the returned pages based on two random keywords separately. After that, the algorithm gets rid of the concrete information, leaving behind the framework labels only like <table> and  in the HTML source codes. Then, the algorithm compares the elements of the two keywords' framework sets to calculate the difference rate. If the rate is more than 90% (an adjustable parameter), the algorithm considers the framework (each of the two frameworks) is what the current search engine is using. Otherwise, the algorithm re-selects two keywords to do the same as above.

Let $\alpha$ and $\beta$ be two randomly selected keywords in Test Words Set at some point T, $P^{\alpha}$ and $P^{\beta}$ the corresponding page sets returned by search engine, $F^{\alpha}$ and $F^{\beta}$ the corresponding framework element set whose elements are got rid of from $P^{\alpha}$ and $P^{\beta}$ . $F^{\alpha}$ .size and $F^{\beta}$ .size are respectively the number of elements in framework element sets $F^{\alpha}$ and $F^{\beta}$ . $F_i^{\alpha}$ and $F_i^{\beta}$ are respectively the element i in corresponding framework element sets. Function WordRandom () randomly chooses one keyword from Text Words set for testing. Function FrameExtract(x) extracts the framework element set from web page x and then returns it. Function ElementEqual(x, y) returns True which represents that elements x and y are the same. DifferCount represents the number of the different elements $F^{\alpha}$ and $F^{\beta}$ have. The concrete algorithm shows below:

**Do**

$\alpha$ = WordRandom(); $\beta$ = WordRandom();

$F^{\alpha}$ = FrameExtract ( $P^{\alpha}$ ); $F^{\beta}$ = FrameExtract ( $P^{\beta}$ );

**For** i = 1 **to** $F^{\alpha}$ .size **Step** 1

**If NOT** ElementEqual ( $F_i^{\alpha}$ , $F_i^{\beta}$ )

DifferCount++;

**End**

**While** (DifferCount/ $F^{\alpha}$ .size < 0.9 **And** $\alpha$ **Is Not** $\beta$ )

*2) Progressive Search Algorithm in Meta-Search Engine*

Taking the fact into consideration that users are used to going on to do progressive search in the wake of their acquiring corresponding information when searching information, we decide to use the computer to simulate this kind of progressive action of human beings in order to acquire more information about the social network. The algorithm is defined as below:

We construct a subset $SubSet_{k_i}$ for every keyword $K_i$ in the keywords set. $K_i$ .RecordSet contains the returned records when we search keyword $K_i$ in the search engine. Then we extract the common characteristic words from the effective data set $K_i$ .RecordSet into $K_i$ .AttributeSet, whose attribute $K_i$ .AttributeSet.size represents the number of elements $K_i$ .AttributeSet has. Function PageSearch ( $x$ ) takes keyword x as its parameter, returning the record set $K_i$ .RecordSet when we search x in the search engine, and function ContentMining(x, y) takes keyword x and its returning record set y as the parameters, returning the common characteristic set of keyword x, namely

$$\begin{cases} y = x.\text{RecordSet} \\ x.\text{RecordSet} = \text{PageSearch}(x) \\ x.\text{AttributeSet} = \text{ContentMining}(x, y) \end{cases}$$

Then we use the software to get every element $K_i$ .AttributeSet[j] from the common characteristic set $K_i$ .AttributeSet searched in the search engine to get the record set $K_i$ .AttributeSet[j].RecordSet for every element in set $K_i$ .AttributeSet. After that, we use the software ROST CM to extract the effective field of $K_i$ .AttributeSet [j].RecordSet to get $K_i$ .AttributeSet[j].AttributeSet which will be combined later with $SubSet_{k_i}$ to achieve the generalized effective data set based on the original keyword. The process is described in the following algorithm.

**For** j = 1 **to** $K_i$ .AttributeSet.size **Step** 1

$K_i$ .AttributeSet[j].RecordSet = PageSearch ( $K_i$ .AttributeSet[j]);

$K_i$ .AttributeSet[j].AttributeSet = ContentMining ( $K_i$ .AttributeSet[j], $K_i$ .AttributeSet[j].RecordSet);

$SubSet_{k_i}$ = $SubSet_{k_i} \bigcup K_i$ .AttributeSet[j].AttributeSet;

**End**

## IV. CASE STUDY

In order to test the difference of the coverage rate between the meta-search engine and a single search engine and validate the effectiveness of the two algorithms above, we design the following three experiments.

### *A. Comparison of Coverage Rate between Google and Meta-Search Engine*

Making use of the cross-page framework adaptive algorithm and the progressive search algorithm in the meta-search engine, we used the ROST Content Mining software to do data acquisition and semantic analysis based on the Chinese-born American physicist, Tsung-Dao Lee, respectively in Google and the meta-search engine. If one scientist appears with Tsung-Dao Lee in one same web page, we add one to the number of their relationship frequency. After checking all the pages related with him, we got all the relationships based on Tsung-Dao Lee. Then we picked out the first 60 high-frequency semantic relationships to form the graph of his social network respectively. The graph shows in Fig.3 and Fig.4

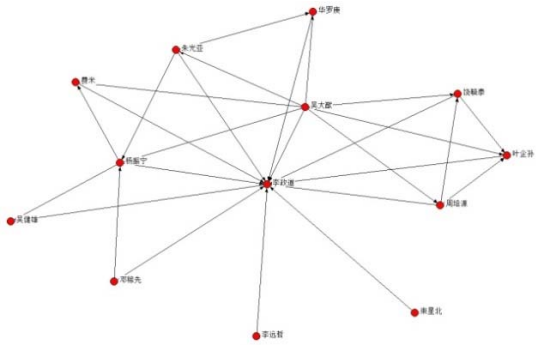

Figure 3. Social Network of Tsung-Dao Lee with Google

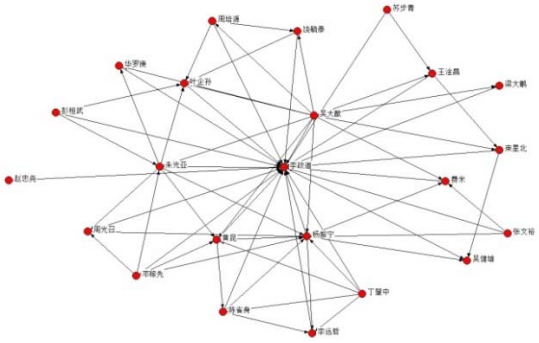

Figure 4. Social Network of Tsung-Dao Lee with Meta-Search Engine

From the above two figures, we get to know that Google covers less effective information than the meta-search engine. The reasons are listed below: firstly, the Google search engine which combines information all around the world is not good at classifying information natively in China, which causes its

application scope relatively narrow compared with the meta-search engine; secondly, the Google search engine returns many Chinese dead links as the result of not often updating the Chinese websites, which surely makes it unable to communicate information effectively; thirdly, Google does its grade division of the pages based on the number of links pages have, on the assumption that the more links one page has, the more useful it is. However, Chinese proper nouns are usually discrete because of their unique meaning, which makes the meta-search engine more useful in Chinese information searching than Google. So, it is meaningful to use the meta-search engine based on the four search engines in our software to acquire the information we need.

### *B. Dynamic Changes in Real-Time Network Observed Through Meta-Search Engine*

Through the cross-page framework adaptive algorithm and the progressive search algorithm in meta-search engine, we made use of the ROST Content Mining System to do the real-time monitor on the social network of Yimou Zhang. Through everyday monitoring, if one word co-occurred with "Yimou Zhang", the relationship frequency between the two words was added by one. After that, we used the first 40 high-frequency semantic relationships to form his semantic graph. We observed the dynamic changes of his semantic graph from 03/01/2009 to 03/12/2009. The graph shows in Fig.5 and Fig.6.

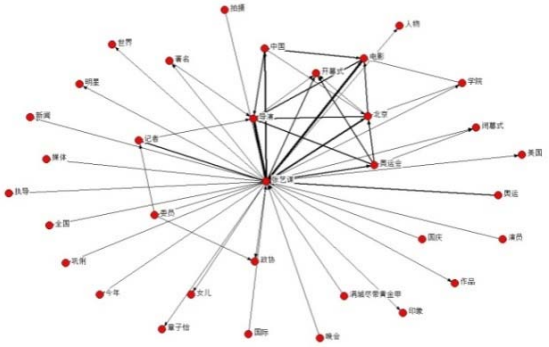

Figure 5. Semantic Graph of Yimou Zhang in 03/01/2009

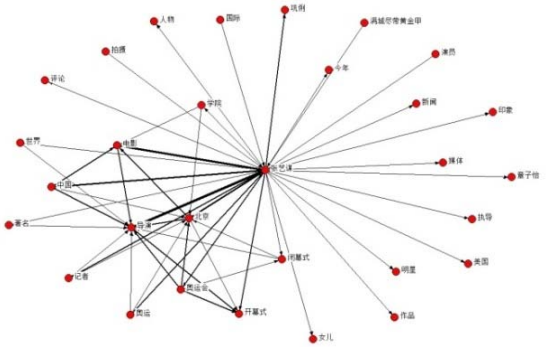

Figure 6. Semantic Graph of Yimou Zhang in 03/12/2009

After comparison and analysis, we get the following conclusions: firstly, the frequencies of "journalist", "news", "committeeman" and "committee" are rising rapidly which is in accordance with the current political affair that Yimou Zhang is attending the Chinese People's Political Consultative Conference and the National People's Congress of the People's Republic of China. This dynamic change of his semantic graph reflects the accuracy of the real-time monitor based on the meta-search engine; secondly, some core words such as "director" and "movie" are always around "Yimou Zhang", showing that this kind of keywords like "Yimou Zhang" have some core attributes that change little; thirdly, the frequency of words like "shoot" and "actor" is decreasing, indicating that his recent behavioral activities do not concentrate on shooting films or choosing actors ; fourthly, the 29th Olympics Games held in Beijing which is one of the physical affairs in the highest competitive level all around the world has brought great effect on the national economy and people's living in China. "Olympics" and "opening ceremony" will be the focus points people talk about for a long time; lastly, people's concern degree on films is going to fall down as time goes.

From the analysis above, we get to know that it is both necessary and feasible to monitor the correlative changes of the semantic network as time goes during a period of time to get to know how one's real-time semantic graph changes.

### *C. Deeper Mining on Effective Content Based on Meta-Search Engine*

By using the progressive searsch algorithm in the meta-search engine and the cross-page framework adaptive algorithm, we input the phrase "invoice service" into the searching bar of the ROST Content Mining System which causes to get the abstracts of the first 2049 pages related with the phrase (including 719 pages from Baidu, 667 pages from Google, 100 pages from Youdao, 563 pages from Sougou). Then we use the regular expression to get 284 telephone numbers related with "invoice", on which we do semantic analysis in the ROST Content Mining System. The graph shows in Fig.7.

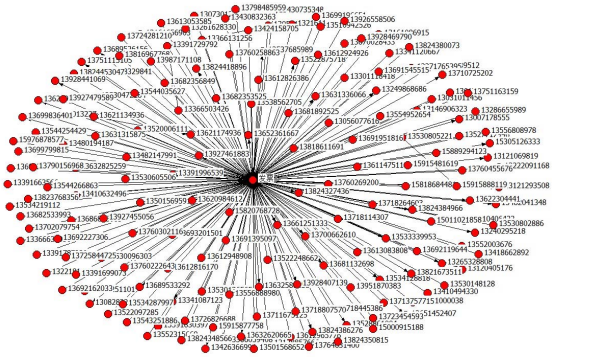

Figure 7. Semantic Graph of Telephone Numbers Based on "invoice"

Then we randomly selected one from the acquired telephone numbers like "13751115105". Based on the progressive search algorithm in meta-search engine, we did the secondary search to get the graph centering the number 13751115105. According to the different frequency of their relationship, we use lines in different bold levels to connect them as Fig.8 shows.

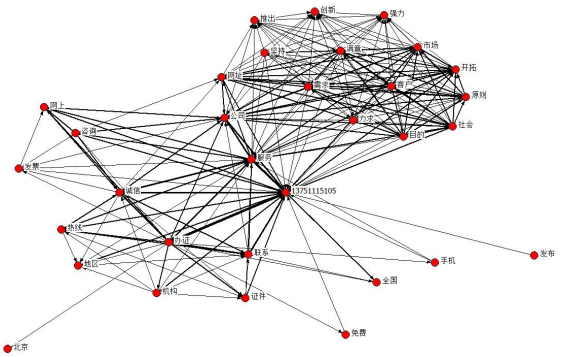

Figure 8. Social Network Based on Telephone Number 13751115105

After respectively extracting the top 8 high-frequency words based on the one-time search and the progressive search in meta-search engine ,we got to find that the results of one-time search is rather vague and general, having no correlative feature representation described in detail. On the other hand, in the progressive search, words with the same properties like "Registration" appeared; appearance of "Honesty" indicated that correlative personnel made such kind of promise; and also the appearance of "Internet" and "Network" reflected the changes that the host of "invoice service" trend to connect their customers through the Internet rather than the paper on streets and lanes in the past. The top 10 high-frequency words are shown in Table.I.

TABLE I
TOP 8 HIGH-FREQUENCY WORD IN BOTH ONE-TIME SEARCH AND PROGRESSIVE SEARCH

| No. | One-Time Search | Progressive Search |
|---|---|---|
| 1 | Service | Registration |
| 2 | Company | Service |
| 3 | Substitute | Contact |
| 4 | Consultation | Honesty |
| 5 | Tax | Company |
| 6 | Beijing | Hot Line |
| 7 | Service Industry | Internet |
| 8 | Advertisement | Network |

Through the secondary progressive search, we could acquire the characteristic words of many telephone numbers, so we can better trace the characteristic information to make a better foundation for mining individual information in a deeper level which will be of great use in searching social characters. We could also make use of progressive search algorithm to find out the difference of the characteristic words of the telephone numbers, so we can know about "invoice service" in a more detailed way.

## V. CONCLUSION

The main contribution of this paper is: firstly, we come up with the new and effective cross-page framework adaptive algorithm and the progressive search algorithm in meta-search engine for better content mining in ROST Content Mining System after analyzing the present situation of content mining based on Google home and abroad; then, we discover the high coverage rate, dynamic time efficiency and deep mining capacity of the social network based on the meta-search engine, which will surely make the analysis method in social network more accurate. In the future, we will do deeper analysis on the current data, and collect more data such as international political news and movie stars to do further dynamic detection and progressive search. At the same time, we will also make research on the formalization of the dynamic subgraph's evolution and further improve the functions of our ROST Content Mining System.

## ACKNOWLEDGMENT

This paper is financially supported by National Natural Science Foundation of China (No. 60803080) and Ministry of Education of the P.R.C. Humanities and Social Science Youth Project (08JC870010), and the National Basic Research 973 Program of China (2007CB310806).

## REFERENCES

[1] Peter Mika. "Ontologies are us: A unified model of social networks and semantics", In Proc.14th International Semantic Web Conference, 2005.

[2] Matsuo Y, Hasida K, Tomobe H, Ishizuka M. "Mining social network of conference participants from the web", IEEE/WIC International Conference, 2003.

[3] Kim Yang Sok,Kang Byeong Ho, Compton Paul, Motoda Hiroshi. "Search engine retrieval of changing information", In Proc. 16th International World Wide Web Conference, 2007.

[4] MATSUO YUTAKA, TOMOBE HIRONORI, HASHIDA KOICHI, NAKASHIMA HIDEYUKI, ISHIZUKA MITSURU. "Social network extraction from the web information [J]", Transactions of the Japanese Society for Artificial Intelligence, 2005, 20:46-56.

[5] Orland Hoeber, Xue Dong Yang. "HotMap: Supporting visual exploration of Web search results [J]", Journal of the American Society for Information Science and Technology,2009,60(1):90.

[6] Jin Yingzi, Ishizuka Mitsuru, Matsuo Yutaka. "Extracting inter-firm networks from the World Wide Web using a general-purpose search engine [J]", Online Information Review,2008,32(2): 196-210.

[7] Y Jin, Y Matsuo and M Ishizuka. "Extracting inter-business relationship from world wide web [J]", Transactions of the Japanese Society for Artificial Intelligence,2007,22(1):48-57.

[8] Gloor Peter A. "Coolhunting for trends on the Web", In Proc. 2007 International Symposium on Collaborative Technologies and Systems, 2007.

[9] Geleijnse Gijs, Korst Jan. "Creating a dead poets society: Extracting a social network of historical persons from the Web", In Proc.1 6th International Semantic Web Conference, 2007.

[10] Lin Ming-Shun, Chen Hsin-Hsi. "Labeling categories and relationships in an evolving social network", In Proc. 30th Annual European Conference on Information Retrieval,2008.

[11] Matsuo Yutaka, Mori Junichiro, Hamasaki Masahiro. "POLYPHONET: An advanced social network extraction system from the web", In Proc.15th International Conference on World Wide Web, 2006.

[12] Peter Mika. "Flink: Semantic web technology for the extraction and analysis of social networks [J]", Journal of Web Semantics, 2005,3(2).

[13] Junichiro Mori, Takumi Tsujishita, Yutaka Matsuo, Mitsuru Ishizuka. "Extracting relations in social networks from the web using similarity between collective contexts". In Proc.15th International Semantic Web Conference, 2006.

[14] Shenyang，ROST Software and Tools Download List http://hi.baidu.com/whusoft/blog/item/6259de2f9e7a2c3c1f3089f9.html. 2009-1-11/2009-3-20