The Social Network Mining of BBS

Lixin Zhou
School of Software and Microelectronics, Peking University, Beijing, China
Zhoulx158@sina.com

Jichang Ding, Yi Wang, Bowen Cheng, Fei Cao
School of Software and Microelectronics, Peking University, Beijing, China
dingjichang@pku.edu.cn, {yifeiyangx, shanshui_66}@163.com, cf_caofei@sohu.com

***Abstract*—Because of the rich and varied resources in internet and social entertainment communities, more and more people or groups are spending their time on discovering the social relationship between others, or finding the core figures for business, political or security reason. In this paper, we build up an SNM solution by analyzing users' static attribute and interaction of BBS to mine what kinds of attitudes they show to others and what kind of relationship between them using data mining technology, to establish users' social network, and to find out the key figures in networks.**

***Index Terms*—Social Network, Data Mining, Core Figure, Six-degree Separation**

I. INTRODUCTION

The rapid development of information technology makes the Internet play an increasingly important role in our lives. Because of its rich and varied resources, the Internet is gradually infiltrated into every aspect of our lives. More and more people would like to spend their time on the Internet especially on some kinds of large social entertainment community. And the more they communicate with each other, the more likely the relationship between them is getting closer.

When one organizational leader is asked to interpret how he/she get the key information that lead to the success of one project, it is not only related to his/her knowledge, but also he/she can effectively find useful information though all of these factors are related to his/her own experience.

Information acknowledgement has made an important but often overlooked part is how a person through his relationships to obtain necessary information. A present research shows that people tend to seek help from friends or colleagues 5 times more than other information resources such as databases or documents. What is more, Internets play a very important role in the procedure of dissemination of knowledge.

Social network analysis is a method of inquiry that focuses on relationships between subjects such as individuals, organizations or nation states. However, it has also been applied to analysis relationships between objects as diverse as the Internet, scientific papers, organisms, and molecules.

Social network analysis has emerged as a set of methods for the analysis of social structures, methods which are specifically geared towards an investigation of the relational aspects of these structures. The use of these methods, therefore, depends on the availability of relational rather than attribute data.

Social network analysis is a method of inquiry that focuses on relationships between subjects such as individuals, organizations or nation states[1][2][3]. If we already had an existing exchange platform for organizations to analyze social Internet to understand the internal and external association of an organization, it will be particularly helpful in enhancing the collaboration, innovating knowledge and disseminating knowledge.

Social network analysis is focused on uncovering the patterning of people's interaction. Network analysis is based on the intuitive notion that these patterns are important features of the lives of the individuals who display them. Network analysts believe that how an individual lives depends in large part on how that individual is tied into the larger web of social connections. Many believe, moreover, that the success or failure of societies and organizations often depends on the patterning of their internal structure. From the outset, the network approach to the study of behavior has involved two commitments: (1) it is guided by formal theory organized in mathematical terms, and (2) it is grounded in the systematic analysis of empirical data. It was not until the 1970s, therefore--when modern discrete combinatorics (particularly graph theory) experienced rapid development and relatively powerful computers became readily available--that the study of social networks really began to take off as an interdisciplinary specialty. Since then its growth has been rapid. It has found important applications in organizational behavior, inter-organizational relations, the spread of contagious diseases, mental health, social support, the diffusion of information and animal social organization.[4]

Therefore, our goal is to build a "social network mining system based on the BBS community". In this system, we will extract personal information and the interaction information between them, doing some data analysis processing, judging the relationship between people and establishing organizations in Internets, to get the organization's internal and external features.

In this paper, we analyze every useful message during their interaction through BBS to mine what kinds of attitudes they show to others, and also get some personal information using data mining technology, to establish the social network, to find out the key figures who are the more active or important BBS users than others. We do some research works through the networks and these key figures.

Section 1 is an introduction, Section 2 gives out the solution overview of SNM, Section 3 described the algorithm implementation of SNM, Section 4 described implementation and experiments, Section 5 described the future work, and Section 6 is the conclusion.

II. Solution Overview of SNM

In recent years, the SNS websites and social network research has advanced significantly, but there still many difficult in mining the BBS, especially using for the BBS evaluation system. In tradition evaluation system, to find out the key figures of the website, to discover the social network, to make a decision who will be the best person to publish an advertisement or notes, to make a choice whether to invested the community, we can only used the statistics from network administrator. Just like: new posts (daily/monthly/yearly), new repliers (daily/monthly/ yearly), person liveness, reputation, it's very hard to find out the relationship between them, and mostly incorrect.

In order to have a deeper understanding about the BBS user, we build up an SNM system. In this solution, we using data mining algorithm on social network to analyze the relationship between users and establish the social network of the BBS for deeper analysis. To address these concerns, we adopted the solution illustrated in Fig. 1.

Firstly, the raw information from BBS are crawled by the web crawler component and gathered into temp files.

Secondly, the pre-treatment component process the raw information from the temp files, which are saved into database after pre-converted.

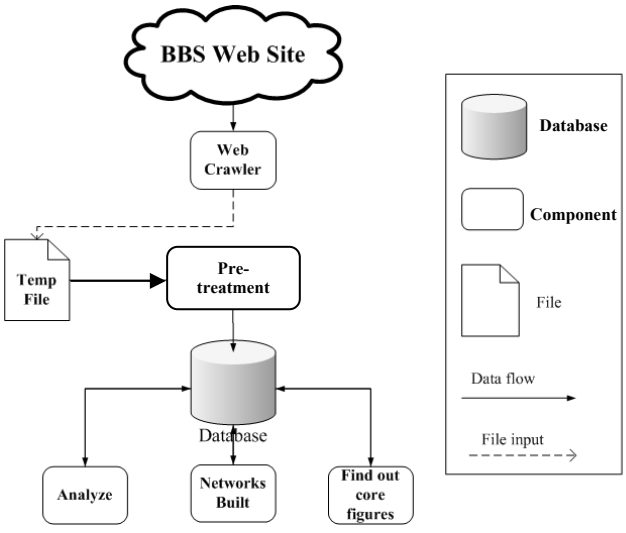

Fig. 1. The SNM architecture

Thirdly, the analysis component analyzes the users' static attribute and interaction between users, and the relationship score will be given.

Fourthly, the networks built component create the social networks for the giving groups, members or crowds.

Fifthly, the find out core figures component is used to find out the core figures from the social network built.

III. Algorithm Implementation of SNM

A. Construction of Data Sources

Construction of data sources include the following processes:

1) Downloading pages: a Cookie supported system should be deployed, which will simulate the steps to enter the BBS system. And then All the Cookies should be saved.

For each HTTP request, Cookie information should be added in, thus we can download the response pages, including personnel information pages and pages on the theme.

2) Mapping data into XHTML: we can use the class library which performs the same functions.

3) Finding the key data: whether in Web pages or the source of XHTML view, most information is the information that we do not care. In this process, we need to filter the unrelated data.

4) Mapping data into XML. Our next task is to find a particular region in the XML tree, thus we can get the related data without the need to care about external information. For a more complex extraction situation, we may need to find several examples of these regions in a single page.

5) Merging results and processing data: the system will read the node in a XML tree and convert it into required format, and then it will insert the converted data into the corresponding tables in database.

B. Analyze the Static Attribute Information of Personnel

Personnel information including the following: UID, registration date, last visit, final page visiting, on-line time, nickname, user group, posting level, reading permission, integration, prestige, money, posts, average posts per day, cream posts, sex, hometown, birthday.

In this process, we try to find some available rue to derive the similarity between two people directly. However, the result will be influenced by a particular platform. We use classified technology according to the messages we already have, refine some useful rules, and then according to the refined rules, following algorithms to analyze the similarity between one person and others:

1) "Mining-frequency".

2) "Impact factor": Assume the personnel information of people A and B contains m items, each score of item similar is s1, s2, s3…sm, and each impact factor of

relationship is p1, p2, p3…pm, so the total similar score of people A and B is:

$$\text{Total}_{a \to b} = s_1 \times p_1 + s_1 \times p_2 + \cdots + s_m \times p_m = \sum_{k=1}^{m} (s_k \times p_k)$$

The more similar two persons' information are, the more probability they will become friends. And based on the information mentioned above, we will analyze the similarity between two persons as a potential factor of judging the relationship between two persons.

C. Analyze the Members' Interaction

Content information may include the following: the author, publish time, theme and content.

If two persons frequently interact with each other, the level of their friendly or hostile may increase. We analyze each word of posts at the same theme according to their response to judge the positive or negative meaning. It will become an important factor to determine the relationship between them.

As for a theme, it is define as a word vector:

Theme ::= [word [,word]$^{0..*}$]

As for a text block of content, it is segmented into Chinese words, and also defined as a word vector:

Content ::= [word[,word]$^{0..*}$]

For each word, it is defined as a positive, negative or others by HowNet[5] semantic dictionary.

We select a theme as beginning, with word vector matching algorithm, analyze all words of each sentence to determine a person's attitude on the subject (positive or negative), thus we will get to know the friendly or hostile attitude between two persons on the theme. Finally after integrating all topics, we can determine the degree of their relationship in friendly or hostile way.

D. Establish the network organizations

In order to find network organizations for BBS data, we have defined a set of organizational rules for a network organization at first. The organizational rules have been stored into a system database for advanced using and the rules can be adjusted for specific targets. In order to find the organization's more internal and external features, we need to establish a network of organizations, for further analysis of the object.

Combining the first two parts' results and accordance with the rules established in advance, the relationship between any two persons could be figured out. Then we can set a threshold of intimacy and choose a certain target as a start point, and groups the points who's intimate level with is more than the threshold, which will form a social network with "weight". The other nodes can be done followed by this way to expand the network. In order to prevent the continuous network expansion, the threshold could be auto-adjusted based on the network si

The network organizations established were showed in a social network graph. A graph is composed of nodes (or users or points) connected by edges (or relations or ties). A graph may represent a single type of relations among the users (simplex), or more than one kind of relation (multiplex)

E. Find out the Core Figure

The core figure may be the most energy or respected members in the organization, his activity may have a significant impact to organization [6]. For the constructed "weight" network, we can calculate the shortest path between the nodes easily. According to the "six-degree separation" theorem, SPLINE algorithm and center-based for mining network sub-groups, we can find out who is the core figure.

The "six-degree separation" theorem describes the nature of network connectivity: The linkers between any two people in any society, only four intermediaries are needed. In other words, any two people on the earth want to know each other, and only jumping 5 steps is enough, also known as the "Tianya 5 neighbor". The theorem came out under the ideal assumption conditions, also the researchers made a large number of social experiments to prove the theorem.

Therefore, for an effective society, there will be not more than four intermediaries to avoid the low efficiency communication, and the shortest path algorithm SPLINE (Shortest Path based on Link weight) was based on the "six-degree separation" theorem. The point is:

1) The frequency nodes via e-mail will be set as the "weight" value between any two nodes;

2) Using Dijistra algorithm to find the shortest path between any two nodes;

3) Pruning, retain the shortest paths which are less than 6.

Basis on SPLINE algorithm, the key points of KMM algorithm to mining the core figure in network groups are:

1) Establish the network while according to the internal communications;

2) The level of clustering approach to mining organization;

3) Using SPLINE and center algorithm to digging in the core figure;

4) Verification of the validity of the algorithm through the experiment, the accurately forecast core rate is 91.2%.

F. The Impact to Organization while Core Figures Mobilize

The core figures' mobilization will have an impact on the organizational structure. Analysis the structure's change while remove a core figure from the organization; analysis the difference while exchange a central figure location between an organization and another; to find if a core figure from other organization join in, whether the two organizations will be merger into one or something else changed?

Assume that the core figures mobilize, observed the changes in the organizational structure.

1) Removed a core figure from the organization; and then re-build network, find out the core figure, analyze the change of overall numbers and the core figure

2) Exchange a central figure location between an organization and another; retaining the comment relationship between the two people and delete the uncomment relationship or use of replacement threshold; and then re-build network, find out the core figure, analyze the change of overall numbers and the core figure

3) A core figure from other organization joined in; establish the link between other organizations and the core figures, the intimate level is the threshold for larger network (usually more); then re-build network below the threshold from a smaller network (usually is small), find out the core figure, analyze the change of overall numbers and the core figure.

IV. Implementation And Experiments

A. The Hardware Environment

We use a personal PC as the system server, which has the following configurations: Intel Dual Core Processor 2.66GHz, 2G Memory, 160GB*SATA Hard disk. The computers are connected to Internet by local network in campus.

The system adopted client – server style. Client also run at PC desktop computer.

B. The Software Environment

In the software environment, the OS is Microsoft Windows Vista, database system is Microsoft SQL Server. Use the C # programming language under .NET Platform to build a "social network mining system based on the BBS community".

C. Major Software Components

The first software component is the web crawler. The web crawler browses the BBS in our campus in an automated manner, providing up-to-date BBS data. The web crawler are used to create a copy of all the visited pages for later processing and social network analyzing, and used to gather specific types of information from BBS for further social network analyzing.

The second component is the pretreatment component. The role of the pretreatment module of our social network mining system is to analyze the structure of the HTML page and to find fields that are relevant to the domain theme. The proper (semantic) way to give the label of a HTML field is the use of the HTML element label, whose attribute must correspond to the name attribute of HTML field. Unfortunately, this tag maybe wrongly used by Web developers and we thus have to resort to other contextual information, such as the name or id attribute, or the text appearing before the field (in the source code). An alternative is to use the graphical lay out of the HTML, and rules indicating where the label of a field most often lies relative to the field.

Structural analysis of an HTML page is as following steps :

Input: URL of a HTML page.
Output: List of fields with theme annotations.

(a) Retrieve the Web page at the given URL from database.

(b) For each HTML field:

(i) Gather all text of the textual content : name and id attributes, content of the corresponding label element, text appearing in the field.

(ii) Segment the text into Chinese words.

(iii) Remove stop-words.

(iv) Check whether any resulting word matches a theme related words as given by HowNet[4].

(v) Annotate the fields with matching themes, with the associated confidence as the probability that this field represents this theme.

The Third component is the analysis component, which analyze the users' static attributes and interaction between others, and calculate the relationship score between users of BBS.

The users' static attributes are extracted from the text data processed by the pretreatment component. The theme word vectors and content word vectors are also created from the data processed by the pretreatment component. Then, we use vector matching algorithm to calculate the relationship score between users of BBS, and calculate a user's attitude score on the theme.

Relationship Score

$$= \overrightarrow{word-of-user1} \times \overrightarrow{word-of-user2}$$

Attitude Score on the theme

$$= \overrightarrow{word-of-user1} \times \overrightarrow{word-of-theme}$$

The fourth component is the network built component, which create the social networks for the giving groups, members or crowds. We use the theme word vectors, users' content word vectors, users' static attributes, group attributes and organizational rules to find groups, members of a groups, and members' relations, then social networks are to be established.

The fifth component is the component for finding out core figures, which is used to find out the core figures from the social network built. By the networks established in the network built component, the core figures are to be found with Graphic Analysis Algorithms[3].

An prototype of social network mining system have been developed and built. The five software component can run on Windows Vista correctly. The experiment results obtained by the prototype system are described and discussed in the following Section D.

Social network analysis can be done by the social network mining system developed. The social network

perspective emphasizes multiple levels of analysis. Differences among users are traced to the constraints and opportunities that arise from how they are embedded in networks; the structure and behavior of networks grounded in, and enacted by local interactions among users.

D. Experiment Results

The test data was collected from BBS website in School of Software and Microelectronics, Peking University.

We have collected 5964 user information HTML pages and 5861 communication information pages from the campus BBS web site. After extracted the users' attributes, parsed analyzed the HTML pages, information of 5964 users and 27840 posts field have been obtained. There are 6 groups which has more than 3 persons mining from the BBS data, whose core figures have been found.

The fundamental data structure is one that leads us to compare how users are similar or dissimilar to each other across attributes (by comparing rows). Or, perhaps more commonly, we examine how variables are similar or dissimilar to each other in their distributions across users (by comparing or correlating columns).

"Network" data (in their purest form) consist of a square array of measurements. The rows of the array are the cases, or subjects, or observations. The columns of the array are -- and note the key difference from conventional data -- the same set of cases, subjects, or observations. In each cell of the array describes a relationship between the users.

We emphasized that the social network perspective leads us to focus our attention on the relations between users, more than on the attributes of users. This approach often results in data that have a different "structure" in which both rows and columns refer to the same users, and the cells report information on one variable that describes variation (in the case of the example below, simple presence of absence of a tie) in the relations between each pair of users.

Fig.2 is User Information pages parsed results. There are eight fields in the table including user id, user name, user last access date, user last post date, access pages etc.

These user attributes information were extracted from 5964 user information HTML pages. 5964 users were extracted.

Fig.3 is HTML pages parsed results. There are eight fields in the table including page id, title, content, tid, uid and date published etc. These text interaction information between users were extracted from 5861 communication information pages on our campus BBS website.

The menu item in Fig.3 high lighted is Web Parser, which means the posts information are to be analyzed and extracted from communication information HTML pages obtained from BBS website. 27840 posts were extracted.

Your goal is to simulate the usual appearance of papers in a Journal of the Academy Publisher. We are requesting that you follow these guidelines as closely as possible

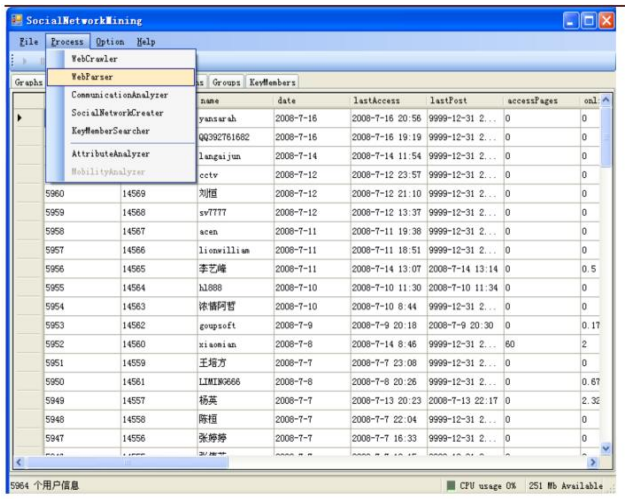

Fig. 2 User Information Parsed Result

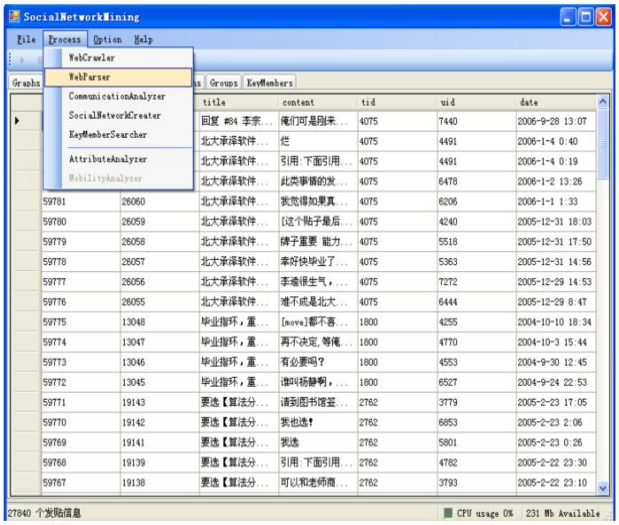

Fig. 3 HTML Pages Parsed Result

These materials aim to help beginners to appreciate and use social network analysis in their own work and pursue further developments on their own.

With the social network mining system we developed, after users' personal information and the interaction information between them extracted from HTML pages on BBS website, we can do some data analysis, judge the relationship between people, establish organizations in Internets, and get the organization's internal and external features. Fig.4 to Fig.7 are some analysis results mined from BBS HTML pages.

***Full network methods*[3]** require that we collect information about each user's ties with all other users. In essence, this approach is taking a census of ties in a population of users -- rather than a sample.

Full network data allows for very powerful descriptions and analyses of social structures. Unfortunately, full network data can also be very expensive and difficult to collect. Obtaining data from every member of a population, and having every member rank or rate every other member can be very challenging tasks in any but the smallest groups.

In this paper, we use Full network methods in social network analysis.

Fig.4 shows the analysis results of relations between users. There are three fields in the table including user1 name, user2 name and relation between user1 and user2. The value in the relation column is 1, means the user1 and user2 has relation in this row.

13025 relations information between users were obtained.

The menu item in Fig.4 high lighted is Communication Analysis, which means the communication relations between users are to be analyzed and extracted from communication information HTML pages obtained from BBS website. If one user answered another user's post, there is a relation between the two users.

Fig.5 shows the analysis results of groups on BBS website. There are two fields in the table including user name and Group ID, named gid. There are six group information showed in Fig.5. A record in a row in the table means the user belongs to the group with the gid.

The menu item in Fig.5 high lighted is Social Network Creator, which means the relations between users are to be analyzed and group information to be mined from HTML pages obtained from BBS website.

Fig.6 shows the analysis results of Key Figures of each group. There are two fields in the table including user name and Group ID, named gid. A record in a row in the table means the user belongs to the group with the gid.

Fig.7 shows the graphic results of networks created. In this result, we find "张莉" (Zhang Li) is the core figure, who has the most ties (relations) with others. This figure was created by using the Graphs tool that is built in our social network mining system. A visual inspection of the diagram with the highlighted node color is much more informative about core figure.

There are only a part of members in a group showed in Fig.7. If all member of a group showed in Fig.7, the density of lines between group members will be very high.

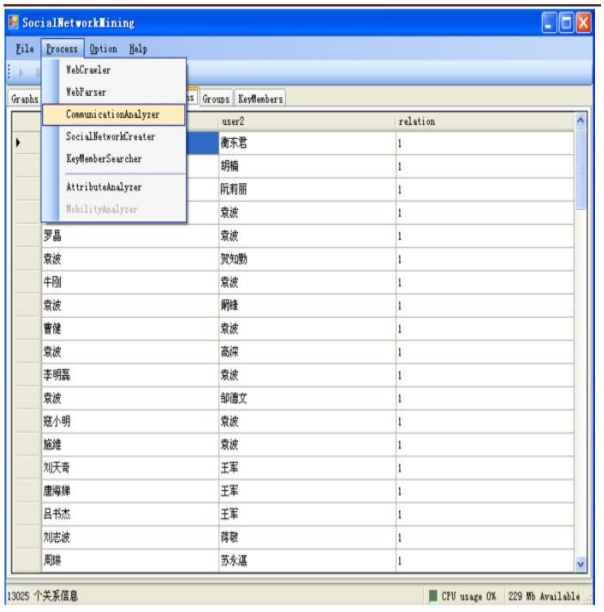

Fig.4 Analysis Results of Relations Between Users

When all member must be showed in a graph, a more efficient plotting algorithm should be adopted.

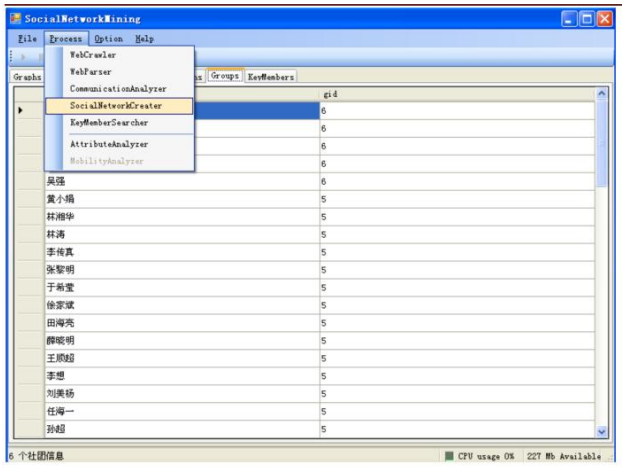

Fig. 5 Group Information

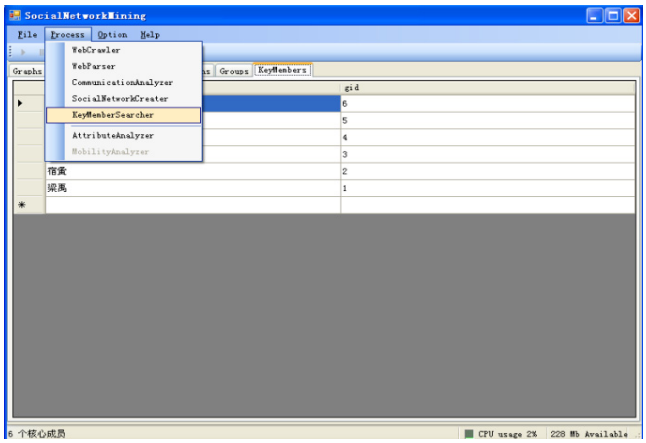

Fig.6 Core Figures Found out

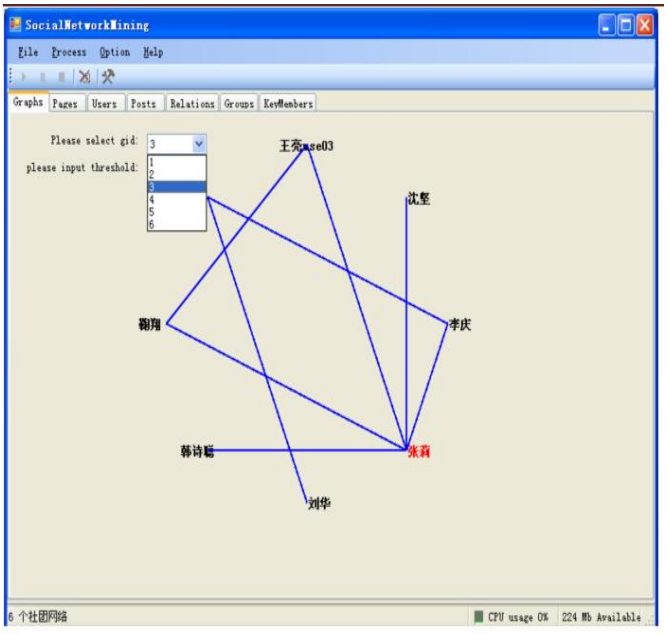

Fig.7 Core Figures Finding out

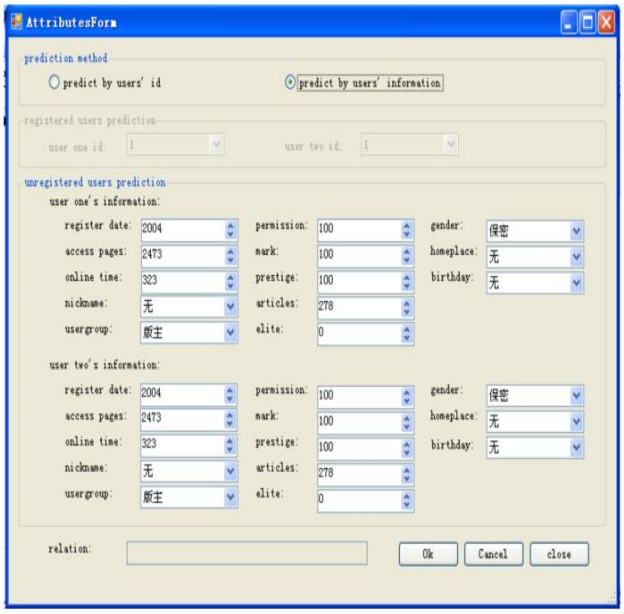

Fig.8 Attribute Analysis Results

Fig.8 shows the user attribute analysis results of networks. The attributes include user register date, access pages, online time, nickname, user group, permission, mark, prestige, articles, elite, gender, home place, and birthday. There are two users' attributes showed in Fig.7. relation can be predicted by users' information.

Differences among individuals in how connected they are can be extremely consequential for understanding their attributes and behavior. More connections often mean that individuals are exposed to more, and more diverse, information. Highly connected individuals may be more influential, and may be more influenced by others.

V. Future Work

Network analysis (or social network analysis) is a set of mathematical methods used in social psychology, sociology, ethology, and anthropology. Network analysis assumes that the way the members of a group can communicate to each other affect some important features of that group (efficiency when performing a task, moral satisfaction, leadership). Network analysis makes use of mathematical tools and concepts that belong to graph theory. A network models a communication group.

In many network studies, all of the ties of a given type among all of the selected nodes are studied -- that is, a census is conducted. But, sometimes different approaches are used (because they are less expensive, or because of a need to generalize) that sample ties. There is also a second kind of sampling of ties that always occurs in network data.

We have obtained some useful results on social network analysis for BBS data mining. In the future, we will increase the scale of the networks and do more kinds of analysis by social network mining.

We also want to add more semantic information in social network analyzing to improve analysis efficiency. We will do more dynamic analysis of social networks.

The most widely used software for social network analysis is UCINET[7] and it is now in its sixth version. It is available for a small license fee for academic use from Analytic Technologies. We will do more useful analysis with UCINET.

VI. Conclusion

In this paper, we build up an SNM solution by analyzing users' static attribute and interaction of BBS to mine what kinds of attitudes they show to others and what kind of relationship between them using data mining technology, to establish users' social network, and to find out the key figures in networks.

We have described the solution to discover the social network, users' relationship, key figures and impaction to the organization on BBS website. Our key idea and contribution is to exploit algorithm and regularities to implementation a solution to discover the potential relationship and social network. The solution and algorithm can be applied to any SNS system that conforms to the implementation conventions, to yield a network in that style.

In this paper, our test data are from Internet and were collected automatically. It is possible to build a large social network.

The social network analyzing and mining tools presented in this paper are useful for organizations to analyze social network to understand the internal and external association of an organization, it will be particularly helpful in enhancing the collaboration, innovating knowledge and disseminating knowledge.

References

[1] Stanley Wasserman, Katherine Faust. Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge: Cambridge University Press, 1994.
[2] Scott, John. Social Network Analysis : A Handbook. Newbury Park, CA : Sage Publications. 1991.
[3] Hanneman, Robert A. and Mark Riddle. Introduction to social network methods. Riverside, CA: University of California, Riverside. 2005.
[4] Lin Freeman: http://www.heinz.cmu.edu/project/INSNA/na_inf.html
[5] http://www.keenage.com/. HowNet Knowledge Database.
[6] Tang Changjie, Liu Wei, Wen Fenlian, Qiao Shaojie. Three probes into the social network analysis and consortium information mining——mining the structure, core and communication behavior of virtual consortium. Journal of Computer Applications, Vol. 26, No. 9, 2006. (in Chinese).
[7] UCINET, Network Analysis program. Available through: http://www.analytictech.com/

Lixin Zhou, obtained Ph.D. in computer science from Zhejiang University, China in 1999. He is currently working as an associate professor in Peking University. His areas of

interests include data mining, information retrieval and natural language processing.

Jichang Ding, obtained B.E. degree in Software Engineering from Huazhong University of Science and Technology, China in 2007. He is now studying Master degree in Peking University. His research areas include data mining, parallel computing and software engineering.

Yi Wang, obtained B.E. degree in Software Engineering from Huazhong University of Science and Technology, China in 2007. She is now studying Master degree in Peking University. Her major field of study are data mining.

Bowen Cheng, is now a graduate student in Peking University. He earned his bachelor's degree in Software Engineering in Huazhong University of Science and Technology, China in 2007. His major field of study is software engineering and data mining.

Fei Cao, obtained B.E. degree in information management from Beijing University of Posts & Telecommunications in 2005. He is currently studying in Peking University. His areas of interests include data mining, data modeling and data analysis.