Social Network Analysis of University Online Forum

Huiqing Niu
College of Economic & Management
North University of China
Taiyuan, China
nnhhqq@163.com

Abstract—To describe the social network in online forum (BBS) of university, social network analysis and data mining method was used to investigate the network relationship of community under the help of UCINET software and ID3 algorithm. Characteristic of community network, formation of opinion leader’s position and the relationship between initiative and other attributes of community members were also explored. On this basis, we proposed some advice in order to solve key problems of the community, for example how to get more better communication in the university online forum, how to make the students more positive, how to make students learn more knowledge here.

Keywords—online forum(BBS), social network analysis, data mining

I. INTRODUCTION

1.1 Overview of concepts

1) Social network

A social network is a social structure made of individuals or organizations called "nodes," which are connected by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, dislike relationships, or relationships of beliefs, knowledge or prestige[1].

Social network has many different styles. Real community is a common kind of social network, and virtual community is at its opposition.

2) Virtual community

The virtual community, which has been constructed by internet, has not only offered a channel for information circulation, but also accumulated the knowledge of the information and formed a huge warehouse of knowledge. Now the virtual community on the internet has been an important platform for knowledge sharing. Computer users have made many social relation networks, and this kind of social network, which is made by internet, is the main foundation of virtual community. The social network in the virtual community is as same as in the real community, it also has the interpersonal relationship features like strong ties and weak ties etc and thus to provide information exchange, knowledge sharing and social support in virtual community.

The members in the virtual community are known each other or unknown each other. But most of them are unknown each other, even they don’t know they are known each other in fact. So the members in the virtual community could give expression their really ideas freely.

Internet provides many tools to construct virtual community. The most popular tool is instant messenger for example MSN, OICQ, and second is online form.

3) Online forum

The online forum, also named BBS, short for Bulletin Board System, has been the most popular tool for university students to communicate and learn each other.

Everyone having the same interesting in the forum could talk about his own ideas, and the other members in the forum could see his idea anywhere and anytime. This is the process of accumulation and aggregation of interpersonal relationship and experience sharing. If the tool is instant messenger, the members who interchange their information must be online in a same time.

The online forum is a more flexible virtual community, even it has no limit of time, let along place.

In conclusion, virtual community is a special kind of social network, and online forum is a special kind of virtual community. Online forum is a most free social network, even it is a most really social network in some degree.

1.2 Overview of methods

1) Social network analysis

Social network analysis views social relationships in terms of network theory consisting of nodes and ties. Nodes are the individual actors within the networks, and ties are the relationships between the actors[2].

The method of social network analysis is a method to study inner construction of social relationship from a perspective of quantization, it can be used to demonstrate and measure the relationship of actors or all kinds of visible and invisible matters, like information or resource, through the relationship[3].

The resulting graph-based structures are often very complex. There may be many kinds of ties between the nodes. Research in a number of academic fields has shown that social networks operate on many levels, from families up to the level of nations, and play a critical role in
d) Network Centralization

Individual network centralities provide insight into the individual's location in the network. The relationship between the centralities of all nodes can reveal much about the overall network structure.

A very centralized network is dominated by one or a few very central nodes. If these nodes are removed or damaged, the network quickly fragments into unconnected subnetworks. A highly central node can become a single point of failure. A network centralized around a well connected hub can fail abruptly if that hub is disabled or removed. Hubs are nodes with high degree and betweenness centrality.

A less centralized network has no single points of failure. It is resilient in the face of many intentional attacks or random failures -- many nodes or links can fail while allowing the remaining nodes to still reach each other over other network paths. Networks of low centralization fail gracefully.

2) Data mining

Social network analysis in the broaden sense, includes all the methods which can solve some problem in social network. Data mining is a method of social network analysis in the broaden sense, also.

The most popular data mining techniques are Artificial neural networks, Decision trees, Genetic algorithms, Nearest neighbor method and memory-based reasoning, Logistic regression, Generalized additive models and Discriminant analysis.

The technique used in the paper is Decision trees. A decision tree is a tree in which every branch is a choice and every leaf is a decision, this kind of trees is often used for information gain in a decision-make process. There are two kinds of decision trees, the classification trees that find the most appropriate class for each case that is analyzed and regression trees used to predict continuous variables.

1.3 Overview of tools

Social network analysis software is used to identify, represent, analyze, visualize, or simulate nodes and edges from various types of input data, including mathematical models of social networks. The output data can be saved in external files. Various input and output file formats exist.

Network analysis tools allow researchers to investigate representations of networks of different size, from small to very large. The various tools provide mathematical and statistical routines that can be applied to the network model.

Visual representations of social networks are important to understand network data and convey the result of the analysis. Visualization is often used as an additional or standalone data analysis method. With respect to visualization, network analysis tools are used to change the layout, colors, size and other properties of the network representation.

UCINET is a social network analysis program developed by Steve Borgatti, Martin Everett and Lin Freeman. The program is distributed by Analytic Technologies. UCINET works in tandem with freeware program called NETDRAW for visualizing networks.
1.4 Research objective

To describe the social network in online forum (BBS) of university, social network analysis and data mining method was used to investigate the network relationship of community under the help of UCINET software and ID3 algorithm.

By learning the community interaction, members in community can set up a learning cooperation relationship. Under the action of community sharing mechanism, the individual knowledge should become a common one in the community and this kind of knowledge could also be structured by specific cooperation.

Nevertheless, the cooperation and communication in virtual community have been influenced by many factors such as the participation motivation and satisfaction of community members etc. And what is the influence on community members’ communication of these factors? How the construction characteristic of virtual community is? How the group members of virtual community and the status of opinion leaders were formed? What’s the influence on the knowledge sharing in community? What is the relationship between initiative and other attributes of community members. These are the topics in our study.

So in this paper, characteristic of community network, formation of opinion leader’s position and the relationship between initiative and other attributes of community members were also explored. On this basis, we proposed some advice in order to solve key problems of the community, for example how to get more better communication in the university online forum, how to make the students are more positive, how to make students learn more knowledge here.

II. MATERIAL AND METHODS

2.1 Material

The subject investigated in this paper, which is based on 48 university students who were organized on Hanquan BBS in North University of China, is the virtual learning community named Object Oriented Programming learning and research group.

Object Oriented Programming is a subject which teach students how to program with Visual Basic, Visual Foxpro, Delphi, and so on. These programming software are all Object Oriented Programming tools because of their same basic programming principle that they look all the things or procedures as Objects.

North University of China is a multidisciplinary university. Many subjects are established in several majors. Object Oriented Programming is such a course. These major who learn the course are Computer Science, Information Management & Information System, Business Administration, and so on. So many students have learned the course. They can talk about the knowledge of the course whatever their major.

In addition, Object Oriented Programming is a software technique in fact. Some students do a lot of experiments and sum up their experience. Some students will have a job to program in the future. So they are enthusiastic to exchange their experience.

The online forum also divided some sub-forum. There is a BBS administrator and there is a BBS moderator in every sub-forum.

The set of university students in online forum is a typical virtual community and it is a social network. The online forum of university has the common characters of social network.

- Nodes: Every student here is a node.
- Ties or Relationships: When a student writes there information on the forum and one of the other students reads the information, there is a tie appeared.
- Objective: Members in the social network have the same values, attitudes, or aspirations: The 48 students had same interesting in the Object Oriented Programming and they had a common goal to learn the subject or the programming language better. This group has offered a new approach for students to study and communicate.

At one hand, the online forum of university has some characters of a common virtual community.

- Virtually: The community is constructed by internet. Their communication through computer and internet, but not through talking face to face. Perhaps these students don’t know each other in life, but they are very good friend here. They needn’t know each other for the topic of a subject.
- Freely: According to the topics about Object Oriented Programming, students can raise questions, discuss their own opinions and exchange ideas to each other freely. Whatever he or she is a good student in fact or not. Everyone could express their views and the other students could accept or oppose his views. This is not only helpful to students’ communication but also to produce innovative thought.

But at the other hand, the online forum of university has some characters of a real community.

- Real: These students of university online forum are from a same university. They written the real information for registering the forum. Otherwise they can not be registered successfully. Since the students don’t know each other, but the administrator has their real information about name, sex, major and so on. It is impossible to achieve in the others virtual community.
- Homogeneity: The members of the university online forum are all the students in a same university, with same age, even with the same or similar major. Their level of knowledge, quality closer. Not every virtual community has such a feature.

2.2 Methods

Social network analysis method will be used to investigate the relationship between members of online forum, depict the construction of community network.
situation and the information flow between these members, explore the relationship of different individuals, discuss the relationship between the network construction and the participation motivation, in order to promote the knowledge exchange and sharing in virtual community.

In this paper, social network analysis method include the visualization of social network, analysis of centralization, analysis of density and data mining.

The data mainly came from an observation by the author and recording by the administrator of the Hanquan BBS in North University of China in 2009. The data include the basic information of every member, the number of times everyone leave message, and the number of times concerned about of each message.

The effective samples were conducted by UCINET software.

III. RESULTS

3.1 The Visualization of Online Forum of University

1) Matrix Approach

Figure 1 is the relationship between community students depicted by matrix. In this matrix, S01 to S48 represent 48 students, \( X_{ij} = 1 \) means that student i and j have the behavior relation. According to the community map, there is a vector line from i to j. \( X_{ij} = 0 \) means that there is no behavior relation between student i and j. So, in the community map, there is no vector line between points i and j. Therefore, we finally get a two-value asymmetric matrix and the community map which is the vector map.

<table>
<thead>
<tr>
<th></th>
<th>S01</th>
<th>S02</th>
<th>S03</th>
<th>S04</th>
<th>S05</th>
<th>S06</th>
<th>S07</th>
<th>S08</th>
<th>S09</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
<th>S13</th>
<th>S14</th>
<th>S15</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S02</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S07</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S08</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1. The Matrix of Relationship among Students of Online Forum in University (Partly)

2) Sociogram Approach

A sociogram is a graphic representation of social links that a person has. It is a graph drawing that plots the structure of interpersonal relations in a group situation. They can diagram the structure and patterns of group interactions. A sociogram can be drawn on the basis of many different criteria: social relations, channels of influence, lines of communication etc.

In this section, in order to analysis the network relationship between community students, we measured the connectivity and figured the social map. As figure 2 shows, orientative arrows represent the connection relationship.

Figure 2 shows the characteristic of the network construction of virtual community in online forum of university as follows:

- **Opinion leaders:**

  There are a number of people, who are respected by the others in the group, they consistently take part in the interaction, own higher popularity, often propose new topic of discussions and direct the interaction of study. They have abundant social intercourse experience and professional knowledge, high performance of solving problems and have strong motivation of study. They are eager to share their experience and help others to study knowledge.

  We call these people “opinion leaders”, and S02, S17, and S48 are all the “opinion leaders”. They publish articles in the community and share knowledge with others. They are the focus of the community and have strong cohesion.

  The opinion leaders are pointed out in the sociogram by bigger squares. If there is a community lack of “opinion leaders”, the social map will be scattered and the net construction relationship will also be scattered.

- **Isolative members**

  We can also find some students whose position are on the edge of community and lack of communication, just like the point S36 and S07, their few connection shows that they were not eager to communicate with other students, seldom publish articles and opinions and made few contribution to community.

  These members are pointed out in the sociogram by up triangles.
3.2 Analysis of Centralization and Density

Centralization is an important index of individual construction position\(^7\). It is usually used to evaluate people’s importance, superiority or privilege of status and social reputation. Generally, the centralization includes: node degree centralization, middle centralization, approach centralization and eigenvector etc, the connectivity are most commonly used.

Connectivity (indegree and outdegree) is commonly used to measure who is the most important people in the virtual study community. It represents the local centralization index of each student in the community. If we focus on the whole net and analysis whether different network lead to different central tendency, the standard point degree central tendency of the whole net can be used to represent the central tendency. The central tendency is more close to 1, the network is more centralized.

1) Analysis of Outdegree and Indegree

The matrix in figure 1 shows the orientation of student relationship, but details of every student are not clear. In this situation, we can analysis according to the indegree, the outdegree and the node degree centralization of every student. The statistical table of connectivity of every student is showed in Figure 3.

Figure 3 shows that different students have different indegrees and outdegrees. Outdegree means the situation that student i has visited other students, and indegree means the situation that other students have visited i. In the matrix of figure 1, outdegree is the sum of cell-grid “1” in the row, and indegree is the sum of cell-grid “1” in the column.

Higher indegree shows the student’s prestige is high and his or her is in the core position. Connectivity is the sum of outdegree and indegree, it is the data to describe how many students have a direct relationship to student i. The standardization of indegree (outdegree) is the ratio between the indegree (outdegree) of one node and the most possible relation number in the network. In Figure 3, the outdegree and indegree of student S17 is 26 and 21, the standardization of outdegree (indegree) is 55.32% and 44.68%, it shows that student S17 pay attention to students and 44.68% students pay attention to student S17. But the outdegree and indegree of student S7 is 26 and 21, the standardization of outdegree and indegree are only 4.26% and 6.38%, it shows that student S7 pay attention to 4.26% of other students and only 6.38% members pay attention to student S7.

<table>
<thead>
<tr>
<th>Node</th>
<th>OutDegree</th>
<th>InDegree</th>
<th>HrmOutDeg</th>
<th>HrmInDeg</th>
</tr>
</thead>
<tbody>
<tr>
<td>S17</td>
<td>26</td>
<td>21</td>
<td>55.32%</td>
<td>44.68%</td>
</tr>
<tr>
<td>S7</td>
<td>26</td>
<td>21</td>
<td>4.26%</td>
<td>6.38%</td>
</tr>
</tbody>
</table>

Table 1. Degree Centralization Measures of students in online forum of university (Partly)

The result of analysis shows that the first 3 students, according to node degree centralization, are S17, S48 and S2. Compared with other students, these students have more connection with others. So they are the “opinion leaders” in the online forum. Among them, half of the students’ indegree is larger than their outdegree and the other half is on the contrary. It demonstrated that half student is focused by more students, and also large outdegree shows that they paid more attention to others. They contribute a lot to the community and get a high status. There are some students whose number of connection relationship is medium. They are the activity of the community such as: S22, S33, S43 and so on. There are 29 activity students. They are pointed out in the sociogram by up circles.

But if we divide in detail, the power and position owned by these students are totally deferent. For example, the outdegree is much higher than the In degree of student S2. It shows that he has paid much attention to others positively but received little and the situation of student S32 and S35 are on the contrary. In addition, some other students have little connection with others, they belong to the inactive student, such as S7 and S36.

2) Analysis of Network Centralization
By the analysis of UCINET, we can conclude that the network centralization (outdegree) is 37.166% and network centralization (indegree) is 34.993% respectively, as we can see in figure 3. The little deference between the two numbers shows that there has no large asymmetry between the network relationships. It has been demonstrated before that the central tendency is closer to 1, the network construction is more focused. According to the central tendency, the whole of network is medium and has no obviously central tendency. So there is no information monopoly by special individuals in the online forum.

3) Analysis of Density
Density is quantitative represent for the overall distribution of social network, is the synergy between social network members in degree. The more members in the social network, the density of the social network is greater. The value of density is actual line between members of the community divided by all possible line. By the software UCINET, we can know density of the online forum of university is 0.2194. The value indicates that the communication in the virtual community is very common. This is a very activity network.

3.3 Data mining application in the university online forum
The research subjective of data mining on university online forum is to find some rule in the relationship between the initiative of members and their attribute. The main attribute influence the initiative of students should be sex, grade, and he or she is BBS moderator or not.

1) The sample data from Hanquan BBS directly.
The sample data from Hanquan BBS directly is shown in table I.

<table>
<thead>
<tr>
<th>No.</th>
<th>Sex</th>
<th>Grade</th>
<th>Outdegree</th>
<th>Indegree</th>
<th>BM</th>
<th>Initiative</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>Girl</td>
<td>High</td>
<td>4</td>
<td>9</td>
<td>No</td>
<td>Passive</td>
</tr>
<tr>
<td>S02</td>
<td>Boy</td>
<td>Middle</td>
<td>3</td>
<td>15</td>
<td>Yes</td>
<td>Active</td>
</tr>
<tr>
<td>S03</td>
<td>Boy</td>
<td>High</td>
<td>4</td>
<td>11</td>
<td>No</td>
<td>Active</td>
</tr>
<tr>
<td>S04</td>
<td>Boy</td>
<td>Low</td>
<td>3</td>
<td>8</td>
<td>No</td>
<td>Passive</td>
</tr>
<tr>
<td>S05</td>
<td>Girl</td>
<td>Middle</td>
<td>1</td>
<td>11</td>
<td>Yes</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S06</td>
<td>Boy</td>
<td>Middle</td>
<td>2</td>
<td>8</td>
<td>No</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S07</td>
<td>Girl</td>
<td>Low</td>
<td>4</td>
<td>15</td>
<td>No</td>
<td>Passive</td>
</tr>
<tr>
<td>S08</td>
<td>Boy</td>
<td>Middle</td>
<td>3</td>
<td>9</td>
<td>No</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S09</td>
<td>Girl</td>
<td>High</td>
<td>2</td>
<td>9</td>
<td>No</td>
<td>Passive</td>
</tr>
<tr>
<td>S10</td>
<td>Boy</td>
<td>Low</td>
<td>1</td>
<td>25</td>
<td>No</td>
<td>Active</td>
</tr>
</tbody>
</table>

2) The sample data settled for data mining.
Assume in the paper is that freshman is in low grade, sophomore and junior is in middle grade, and senior is in high grade. Because sophomores and juniors have similar characters.

The other assume initiative is represent by sum of outdegree and indegree. The highest thirty percent members is active, the middle forty percent members is middle, and the last thirty percent members is passive.

The data settled for data mining is shown in table II.

<table>
<thead>
<tr>
<th>No.</th>
<th>Sex</th>
<th>Grade</th>
<th>BM</th>
<th>Initiative</th>
</tr>
</thead>
<tbody>
<tr>
<td>S01</td>
<td>Girl</td>
<td>High</td>
<td>No</td>
<td>Passive</td>
</tr>
<tr>
<td>S02</td>
<td>Boy</td>
<td>Middle</td>
<td>Yes</td>
<td>Active</td>
</tr>
<tr>
<td>S03</td>
<td>Boy</td>
<td>High</td>
<td>No</td>
<td>Active</td>
</tr>
<tr>
<td>S04</td>
<td>Boy</td>
<td>Middle</td>
<td>No</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S05</td>
<td>Girl</td>
<td>Middle</td>
<td>Yes</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S06</td>
<td>Boy</td>
<td>Middle</td>
<td>No</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S07</td>
<td>Girl</td>
<td>High</td>
<td>No</td>
<td>Passive</td>
</tr>
<tr>
<td>S08</td>
<td>Boy</td>
<td>Middle</td>
<td>No</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S09</td>
<td>Girl</td>
<td>Middle</td>
<td>No</td>
<td>Ordinary</td>
</tr>
<tr>
<td>S10</td>
<td>Boy</td>
<td>Low</td>
<td>No</td>
<td>Active</td>
</tr>
</tbody>
</table>

3)ID3 Algorithm
The technique of data mining used in this paper is Decision tree, and the algorithm of Decision tree used in this paper is ID3 Algorithm. ID3 Algorithm builds a decision tree by starting a top-down search through the given sets to test each attribute at every tree node. The algorithm uses a greedy search, that is, it chooses the best attribute and never reconsiders the choices he made.

a) Main steps
The main steps of the ID3 algorithm are[8]:
- For each attribute in the data table, compute its entropy;
- The current node is the attribute (A) with highest information gain;
- For every value of the attribute A build a subtree; if \(A=value1\) then generate subtree1; if \(A=value2\) then generate subtree2, etc.
- For each subtree, repeat this process from the first step;
- Every time a new node is created in the tree with a variable, that attribute is removed from the variables group;
- The process stops when there are no attributes left.

b) Entropy[9]
In information theory, entropy is a measure of the uncertainty about a source of messages. The more uncertain a receiver is about a source of messages, the more information that receiver will need in order to know what message has been sent.
Given a collection \( S \) of \( c \) outcomes.
Entropy(S) = \( \sum -p(I) \log_2 p(I) \)

Where \( p(I) \) is the proportion of S belonging to class I. S is over c. \( \log_2 \) is log base 2.

That S is not an attribute but the entire sample set.

c) Information gain

How does ID3 decide which attribute is the best? A statistical property, called information gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information is selected. In order to define gain, we first borrow an idea from information theory called entropy. Entropy measures the amount of information in an attribute.

Gain(S, A) is information gain of example set S on attribute A is defined as

\[
\text{Gain}(S, A) = \text{Entropy}(S) - \sum (|S_v| / |S|) \times \text{Entropy}(S_v)
\]

Where:
- \( \Sigma \) is each value v of all possible values of attribute A
- \( S_v \) = subset of S for which attribute A has value v
- \( |S_v| \) = number of elements in \( S_v \)
- \( |S| \) = number of elements in S

In the case of university online forum, the attributes are sex, grade, and BM, short for he or she is BBS moderator or not, short for. They can have the following values:

- Sex = \{boy, girl\}
- Grade = \{high, middle, low\}
- Is BM = \{yes, no\}

Step 1: Find which attribute will be the root node in our decision tree.

The gain is calculated for all attributes:

- Gain(S, Sex) = 0.125
- Gain(S, Grade) = 0.308
- Gain(S, BM) = 0.029

Grade attribute has the highest gain, therefore it is used as the decision attribute in the root node.

Since Grade has three possible values, the root node has three branches (High, Middle, Low). The next question is "what attribute should be tested at the High branch node?"

We have used Grade at the root, we only decide on the remaining two attributes: Sex and BM.

This process goes on until all data is classified perfectly or we run out of attributes.

Then we generated the decision tree by using the entropy and the information gain for each node the tree is split. The result tree is shown in figure 4.

The decision tree can also be expressed in rule format:

IF Grade = low THEN Initiative = active

IF Grade = middle AND Sex = boy THEN Initiative = active

IF Grade = middle AND Sex = girl THEN Initiative = ordinary

IF Grade = high AND BM = yes THEN Initiative = active

IF Grade = high AND BM = no THEN Initiative = passive

In conclusion, the freshman is most active, and the sophomore and junior are ordinary, if he or she isn’t a BBS moderator, the senior is very passive.
IV. DISCUSSION

From what has been discussed above, we may draw these conclusion.

The online forum of university is belongs to mesh construction. It is accordance with the community centrality. The relationship of network is symmetric. The online forum is a successful virtual study community.

We think that the reason why this study community has succeeded as follows:

- Firstly, most members in the community have certain participation. They can actively take part in the study on internet and communicate to each other.
- Secondly, there have some central figures in the community, they guide the operation of the whole community.
- Thirdly, the students understand each other more better than the other social network, because they have a strong homogeneity.

Moreover, we found that community has some disadvantage.

- Firstly, there are some “isolative members”, they have low participative motivation and low satisfaction and loyalty to the community.
- Secondly, the relationship between community members is symmetric and there are few small group in community which lead to a centrality topic.
- Thirdly, members in community have a strong homogeneity that will lead to a uniform role and low community activity.
- At last but not least, there are too much freshman and too few senior. It leads to knowledge unbalance.

Therefore, in order to a continuous development of community, some advices are proposed.

- Add some new member to reduce the most high homogeneity, for example teachers, employee of software.
- The generation of “opinion leader” should be emphasized and transform the “isolative members” into active members, and then, transform the active members into “opinion leaders”.
- The generation of small group should be emphasized because it is advantage to abundant the topics and lead to deeper cooperation.
- Specific methods will be adopted to enhance the enthusiasm of senior, because they are knowledgeable and experienced.

V. SUMMARY

This paper used social networks analysis method and UCINET software, discussed the network relationship in online network of university, obtained some index (Indegree, Outdegree, Centralization and Density) relate to the participation motivation of community members.

This paper also analysis the sample data of university online forum with a Decision Tree method of data mining—ID 3 algorithm. The relationship between initiative and other attributes of community members were explored.

At last this paper discussed and proposed some suggestion for development of online forum of university based on this basis.

Because it is a case research, it is hard to spread the result to other situations. It needs a further research.

ACKNOWLEDGMENT

We want to thank the administrators of Hanquan BBS, the online forum in North University of China, for their participation in the process of information gathering. We also wish to thank the professor Chen Hong, who gave us most guidance and encouragement.

REFERENCES

[8] Monash University, Faculty of Information Technology, CSE5230 Tutorial: The ID3 Decision Tree Algorithm, Semester 2, 2004

429