Network Community Discovery Method based on User Link and Interest

Jianzhuo Yan, Ying Wang, Liying Fang, Hexin Duan, Mengyao Qi
College of Electronic Information and Control Engineering,Beijing University of technology
Beijing, China
E-mail: :wangying8828@sina.com

***Abstract*— In order to enhance the service quality of network community, satisfy the growing demand of community users, this paper analyzes the content of network community from different aspects and provides a comprehensive method to divide the network community user group. On one hand, this method use link analysis techniques to study hyperlink, calculate the number of output links and input links, and then construct the diagram of community user relationship and divide user group. On the other hand, this method analyzes user's interest which is expressed in the published articles and reviews based on support vector machine (SVM) classification. We use clustering to divide group based on the characteristics of the user's interest. A comparative analysis of the different results can get the final result with higher accuracy and reliability. At last, we use software of social network analysis to evaluate the results. This method provides a theoretical basis and technical means for network community application's optimization and personalized services.**

Keywords-cluster; classification; network user group; link analysis; user's interest

I. INTRODUCTION

With the development of Web2.0, Internet has entered a new stage and become the main medium of information exchange and social interaction. According to China Internet information center （CNNIC）released the latest China Internet network development statistics report[1] which shows that Internet users in China has reached 564 million, Internet penetration rate reached 42.1%. The growth in the number of users, promotes innovation and the development of network services and applications. Online community provides a resource-rich, convenient online platform to people, and collects all sorts of interests, hobbies or aspiration of people together. People express feelings and discuss problem by the means of releasing information. In the online community, users are often dominant to perform concern and associated by manual links, or express interest orientation through published articles and reviews. So users create recessive relationship by common hobbies.

Exploring explicit and implicit relations between the online community users, it has a fundamental effect to improve and optimize the network community applications services. At present， the main research work has following several aspects: describe the framework and build the model of online community [2][3]; optimize and improve the network community function, such as search and recommend [4][5];analyze user characteristics and evaluate user's loyalty[6][7].

Compared with other people's work, this paper analyzes online community users ' links and interest, divides user group .It not only analyzes user's characteristics and describes the structure of network community, but also provides richer user's information for optimizing community function. This article combined link analysis with text interested preference analysis, to divide the network community user groups. The paper verifies and supplements each other through the two ways, and increases the accuracy and reliability of the division results. This approach is not longer simply to make link relations as the basis, emphasis the rich text information is very important in the process of network community user group division.

II. PRINCIPLE AND PROCESSION OF ANALYSIS

A. Principle

On one hand, this paper studies the network structure of online community user's link by link analysis method. On the other hand, the paper uses classification and clustering analysis method to analyze online community users 'same interest in the structure of text content, and then we make difference fusion to two results. At last, we can get comprehensive results of online community user groups division. On this basis, we use social network analysis each divided result by evaluate network density, group radius and other indicators.

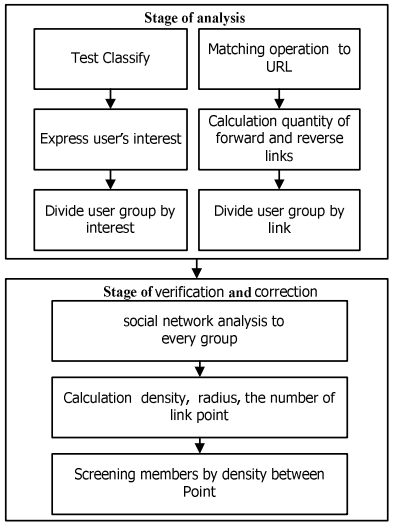

Figure 1. Process flow diagram of analysis method

978-0-7695-5004-6/13 $26.00 © 2013 IEEE
DOI 10.1109/ICCIS.2013.172

B. Process

The concrete implementation steps of this way are as follows:

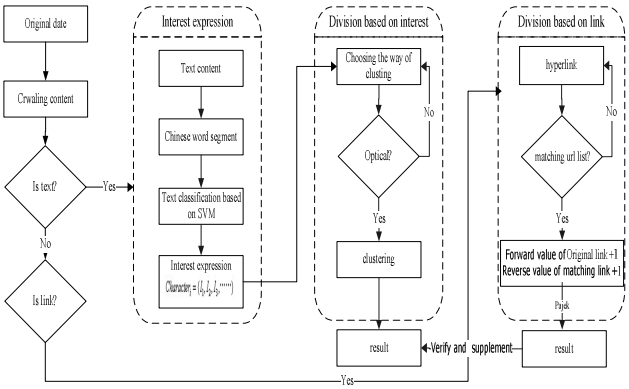

Figure 2. Diagram of concrete implementation steps

Step 1: crawling the content and link of webpage, which included text, discuss, reference links and friendship links. Storing in the text database, link database respectively

Step2: Using IKAnalyzer to realize Chinese word segmentation and then using text classification method is based on SVM to realize text classification

Step3: Analyzing the category of all pages for each user. The category express interest orientation of network community users, which is shown by vector as follows:

$$Character_i = \{I_1, I_2, I_3, \cdots\cdots\}$$

Among the vector, character is the user interest feature set, I_i is interest's Feature weights.

Based on this, network community users can expresses by matrix forms:

$$\begin{bmatrix} I_{11}, I_{12}, I_{13}, \ldots I_{1n} \\ I_{21}, I_{22}, I_{23}, \ldots I_{2n} \\ \ldots \\ I_{m1}, I_{m2}, I_{m3}, \ldots I_{mn} \end{bmatrix}$$

Among of them, n is the number of interest character, m is the number of users.

Step4: Comparing the clustering method from the big amount of data processing capacity, data sequence sensitivity, parameter dependence. Algorithm is based on the division and based on the level and based on the density .Choosing the optical one between them.

Step5: Using the optimal cluster algorithm to divide the network user group. The input is matrix of user's interest.

Step6: Reading sequentially link database orderly, crawling friendship links and reference link circularly, analyzing the frequency of each link appeared in the webpage.

Step7: Pajek provides a complete set of complex networks analysis algorithm. The input is input-degree and output-degree of links. Using Pajek to analyze and display user affinity relationships and group classification clearly.

Step8: Analyzing the result of classification is based on link .and comparing with the result of classification is based on content. And then do difference fusion for two results

Step9: Calculating network group density, radius and other indicators which are based on the result of links and content. To verify the correctness of division's result.

$$\text{Group density } \rho = \frac{2 * N_{line}}{n * (n-1)} \quad (1)$$

N_{line} is the number of line, n is the number of point

$$\text{Radius } D = \max\{d(i,j)\} \quad (2)$$

$d(i,j)$ is the distance between i and j. D is biggest distance.

III. EXPERIMENT

This paper selects Sina.com's blog as the experiment object of division network community user groups, and chooses 1000 bloggers as research objectives from blog

After Chinese word segmentation and text classification, we need to express users' interests. User's interest character includes entertainment, economy, literature, science and technology, healthy, military, and sport.

The paper does contrast on the clustering method from the big amount of data processing capacity, data sequence sensitivity, parameter dependence. Algorithm is based on the division and based on the level and based on the density The result of comparing as follows:

TABLE I. COMPARING RESULT OF CLUSTERING ALGORITHM

	K-means	**DBSCAN**	**Hierarchical**
data processing capacity,	2.25s	2.6s	6.13s
data sequence sensitivity,	Don't sensitivity	Don't sensitivity	sensitivity
parameter dependence	dependence on K	dependence on density	Don't dependence

We can see from the table above, K-means perform well on he big amount of data processing capacity, data sequence sensitivity, but the result of clustering is dependence on K. The paper calculates minimum value of different class samples and the average distance of same class samples to solve problem.

Minimum value of different class samples

$$b(j,i) = \min_{i \le k \le c_i, k \ne j} \left(\frac{1}{n_k} \sum_{p=1}^{n_k} \left\| x_p^{(k)} - x_i^{(j)} \right\|^2 \right) \quad (3)$$

For K,j indicate class , $x_i^{(j)}$ is a sample in class j, n_k is the number of class k.

Average distance of same class samples

$$w(j,i)=\frac{1}{n_j-1}\sum_{q=1,q\neq i}^{n_j}\left\|x_q^{(j)}-x_i^{(j)}\right\|^2 \tag{4}$$

For $x_q^{(j)}$ indicate a sample in class j, $q\neq i$, n_j is the number of class j.

The sum of them is $baw(j,i)=b(j,i)+w(j,i)$, the difference of them is $bsw(j,i)=b(j,i)-w(j,i)$

$$k=\arg\max\{\frac{1}{n}\sum\sum U\}=\arg\max\{\frac{1}{n}\sum\sum\frac{bsw(j,i)}{baw(j,i)}\} \tag{5}$$

So we can get the optimal value of K.

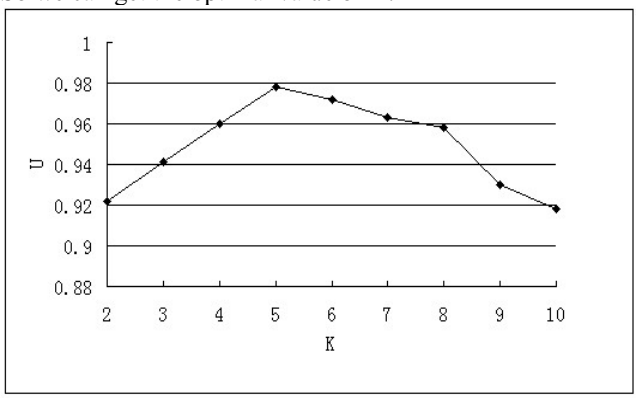

Figure 3. The curve of U and K

From the above figure, we can know five is optimal number of clusters. We use K-means algorithm to cluster. The input is matrix of user's inertest .So we can get the five user group. The result is as follows:

TABLE II. THE USER GROUP DIVISION RESULT BASED ON TEXT

Entertainment	Economy	Literature	IT	Healthy	Sport	Military	Id	Class
12.9	19.35	16.94	9.68	0.81	40.32	0.00	4	1
2.55	16.36	8.73	49.45	0.00	20.0	0.00	7	1
7.41	25.93	20.37	1.85	0.00	42.59	0.00	11	1
0.00	23.81	12.38	31.43	0.00	24.29	0.48	12	1
3.85	30.29	11.54	17.31	0.00	33.65	0.00	14	1
14.71	23.53	16.91	11.03	0.00	30.88	1.47	18	1
4.36	18.12	15.94	12.32	0.00	47.10	0.00	33	1

We can see from the table above, class 1 includes No. 4, 7, 11, 12, 14, and other users. These users have common interests, which are having great interest in sports and economy, and don't pay more attention on health and military information.

The paper divide network community user group by calculating the value of link's input-degree and output-degree. The result as follows:

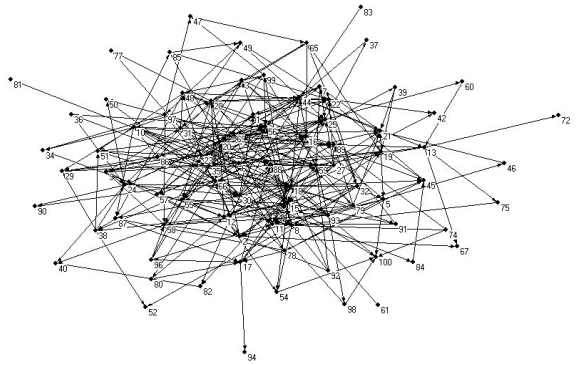

Figure 4. User group division result based on link analysis

After we analyze and compare the two kinds of division result, finding the user group division result based on link analysis, which is lack of No.14, 26, 33,42,4 and other nodes in the result based on the link. The result describe that link relationship does not exist between these users. But from the division results based on the text content, we can know that they have common characteristics and interests. So we need to add the result based on the text content to the result based on the link. And the result based on the link with the text content as follows:

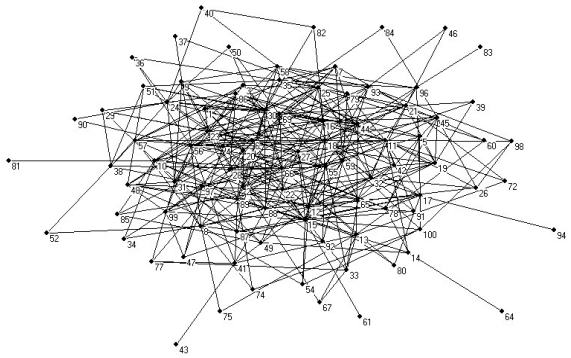

Figure 5. User group division result based on link and text

According to the above result, we can calculate network group density, radius which is based on the result of links and content. They verify the result of division and removal point does not meet requirements. The experiment shows that group density is 0.0435718, radius is 12.Generally, when radius is bigger than 0.01, indicate that the network connection between point and point more closely. The 18th node is the center node of the entire network and the 81st connect with other nodes does not close. It means the users No.81 to connect with other users do not frequently. So we can remove this node.

IV. CONCLUSION

The existing network community user group division method is based on links, but this paper considers implied relationship which is expressed by same hobbies. For example having four users which are A, B, C, D. Among of them, A and B and C have link relationship, so ABC is a

group according to the exciting method. In the fact D and ABC have the common interest. Using the method of this paper we can known ABCD have relationship and should be divide into same group.

In the paper, we use method of social network analysis verify results. At the first time, we compare the result of user group division which is based on link and text content. At the second time, we use indicators which are density, graph diameter, group of radius, the link point number, and center of group to inspect the result of division. The analysis results have high credibility and reliability. This method provide a way to network community understand the user and divide user group ,and provide technology basis for active service of community. So it is very important to optimization of network community and very useful to analyze the structure of network community. The most important is that it provides a better way to discovery network community.

REFERENCE

[1] CNNIC.China Internet Network Development Statistics Report 2013

[2] R. Kumar, J. Novak, A. Tomkins, "Structure and evolution of online social networks", *Proceedings of 12th International Conference on Knowledge Discovery in Data Mining, ACM Press*, 2006, pp. 611-617

[3] Kan Li ,Yin Pang, Avertex Similarity Probability Model for Finding Network Community Structure. *Advances in Knowledge Discovery and Data Mining - 16th Pacific-Asia Conference, PAKDD* 2012 pp456-476

[4] Gao Yan, Fu Li, A method of perfecting the community network 2012 CN102929889A

[5] Huang,LiMing A methods to automatic identify friends 2011 CN1021857872.

[6] Marcelo Maia, Jussara Almeida, Virgílio Almeida, Identifying User Behavior in Online Social Networks, *Proceedings of the first Workshop on Social Network Systems*, 2008

[7] Huang Lailei,, Xia Zhenyou,User. Character and Communication Pattern Detecting on Social Network Site,*I nternational Conference on Engineering Computation* 2009.p.261-p.265