Identifying and Evaluating the Internet Opinion Leader Community Through k-clique Clustering

Jianfang Wang
Department of Computer and Information Technology, Nanyang Normal University, Nanyang, China
Email: jianfangxyz@sina.com.cn

Xiao Jia
Department of Computer Software, Nanyang Normal University, Nanyang, China
Email: xiaojia66@126.com

Longbo Zhang
School of Computer Science and Technology, Shandong University of Technology,Zibo, China
Email: lbzhang@sdut.edu.cn

***Abstract*—With the rapid development of the Internet technology, the Internet has become an important source of information for many acquiring the public sentiment. Opinion leaders play an important role in leading in the direction of the public opinion. In this paper, we drew the community components from the replies of every post in BBS according to the structure of the community in the network, and we came up with a method of extracting the opinion leader community (OLC) based on the hierarchical structure. In this way there were more overlapping appearances among members of the communities. Thus, the relationship between any two communities can be enhanced, which makes it easier to identify the OLC. Then, we analyzed the revolution of the OLC and put forward a time-dividing method of dividing the whole communities into different parts based on the characteristics of the post and the time period and gave the suitable measurement parameter to get the evolution result of the communities. Finally, experiments proved the efficiency of the OLC extracting method and the properties of the OLC revolution were summarized.**

***Index Terms*—Public Sentiment; Opinion Leader; Discovery of the Community; Revolution of the Community**

I. INTRODUCTION

With the rapid development of the network, social computing which is a cross-disciplinary between modern computing technology and social sciences has been more and more important. Social computing, in other words, which is faced with social activities, social processes, social structure, social organization and social function is a computing theory and method. On one side, it researches the applications with computers and information technology in society. On the other hand, it is based on social science knowledge, theory and approaches using the power of computing technology and information technology in order to help people know and research the problems in social science and enhance the efficiency and the level of human social activities.

BBS has been a popular communication platform in which people reflect the phenomenon and the status of current society and show the trend of development of the society[1]. There should find people called opinion leaders who have the ability of leadership and the central role. The opinion leaders are activists that often provide information, opinions, comments, and influence to others in interpersonal communication networks and the intermediary of the formation of the mass effect or the filter links. By analyzing their comments, we can understand of the general trend of each post in advance. Therefore, it is important for us to identify the opinion leader community in the public sentiment.

Besides, it is helpful for people to understand the dynamic law of the whole network by finding the evolution of the opinion leader community in the public sentiment. Therefore, we can achieve optimal network structure, resources, search, and resources, with the basis for such recommendation. However, existing community discovery methods often just extract the parallel relationship between community structures. These communities are either independent of each other or overlap. They do not consider the level relations that may exist between the communities. What is more, it is not reasonable to partition the types of the evolution of the communities. Community change is often complex, it is difficult to describe a community in an accurate way only with increasing, reducing, combining and so on. [1, 2]

This paper is organized as follows. In section 2, the related works are introduced. In section 3, the method of mining the opinion leader community and how to evaluate the evolution of the community are presented. The experimental results and the analysis in this paper are also presented in section 4. Finally, our work of this paper is summarized in the last section.

II. RELATATED WORK

doi:10.4304/jcp.8.9.2284-2289

Among all the methods of mining community, Kemighan-Lin method [3] which is a greedy algorithm based approach that divides the net into two known pieces. It should require knowing the size of the community in advance. The complexity of the algorithm based on Laplace [4] matrix Eigen value and the traditional method based on spectral split resistor network voltage spectrum Wu-Huberman algorithm are low, but they are required to know the number of community networks, but in actual network analysis, the number of communities is often the unknown. Some of these parameters in the methods are not very clear, so they are not convenient to use.

GN algorithm [5] is representative of splitting algorithm; the basic idea is constantly removed from the network side of the largest referral number. GN algorithm has two main shortcomings, namely the lack of definition of the amount of community structure, the second is the need to double counting. Newman [6] is a fast algorithm based on greedy algorithm ideological cohesion, which applies to large-scale complex networks. Newman algorithm's defects are either split or clustering algorithm, the network is divided into separate communities. However, the networks of community are often interrelated and overlapping.

In order to extract reasonable structure of the community, the community discovery method introduces time properties. Toyoda studied the evolution of the Web community [7]. A Web community is constituted by a number of web pages, the web pages are connected by hyperlinks connected to the dense. These pages are usually on the same subject, the changes of the community directly response to changes of the Web topic. Through the respective web pages for 1999,2000,2001,2002 we build the network and extract the archive community, creating community maps, analyzing the evolution of Web communities and the phenomenon of the emergency. In analyzing the evolution of the community, the introduction of a series of metrics, such as growth, stability, novelty, disappearance rate, fusion rate, cleavage rate are used to describe the complex changes in the community [8].

Therefore, we have adopted a community-based level discovery method, and made a number of algorithms to find opinion leaders in the extraction of the community，and analyze the evolution of the network community according to the different nature of the forum posts using different length of time division, and proposed measurement mechanism.

To figure axis labels, use words rather than symbols. Do not label axes only with units. Do not label axes with a ratio of quantities and units. Figure labels should be legible, about 9-point type.

Color figures will be appearing only in online publication. All figures will be black and white graphs in print publication.

III. Parallel Proving Algorithm based on Semi-Extension Rule

In this section, we use level-based community extraction method when extracting opinion leader community. Besides, we come up with some approaches based on LCS and DFS algorithm in mining the opinion leaders.

We believe that a community can be seen as some collection of connected groups in some sense. Group is a fully connected network diagram. Group consisting of k nodes is called the k-groups. If there are k-1 common nodes between two groups, we called them adjacent. If a k-groups can reach another k-groups through a number of adjacent k-groups, we called the two k-groups are connected. K-groups community in the network can be seen as a collection consisting of mutually connected k-groups. We can define k-groups chain as the union of adjacent k-groups, and introduce the k-groups connectivity: two k-groups are connected if they are a part of k-groups chain. We think the k-groups community equals to the connected component of the k-groups.

Clustering all of the k-groups is a necessary condition to find a k-groups community. At first, we should estimate the maximum k from the network according to the degree of each node; then, find the k-groups which contain the certain node; delete the edge related with the node in last step and repeat the second step until finding all of the k-groups. At last let k be k-1 to do the same thing above to find the whole community [9].

The method to cluster the k-groups community is as follows. We define a certain node v, two sets A and B. Set A contains all of the nodes related with the node v. Set B is consist of all the neighbor nodes of set A. For a certain S:

(1) Initialize set A and set B;

(2) Move a node of B to set A, and delete the nodes that are not connected with the nodes in set A;

(3) If |A|<S and |B|=0, or set A and B are the subset of clique, stop computing and return to the second step;

(4) If |A|=S, record the clique, and return to the second step.

In this way, we can cluster k-groups community from node v with certain size.

In addition, when we find opinion leaders in the community we have found above, we should analyze the nodes in the cliques.

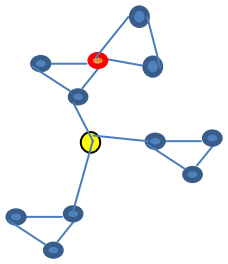

Figure 1. Diagram of Analysis of the Nodes

There are two types of nodes in the community. The ones that are not in any of the cliques and the ones those are not. We should analyze them both.

(1) For the nodes that are in the cliques, the red node, for instance, is an intersection of two cliques and it is more important. However, the blue nodes are not which are less important. In this way, we put forward the method based on LCS algorithm.

(2) For the nodes that are not in the cliques, the yellow node, for example, is the connection of three cliques. It is also more important for us. It can be regard as a opinion leader. Therefore, we come up with an approach based on DFS algorithm.

For two sequences, Xm = x1 x2 x3 ... xm and Yn = y1 y2 y3 ... yn, the algorithm is as follows:

Algorithm1：LCS

Input：str1[len1]，str2[len2]；

Output: Two strings of the longest common subsequence, and its length

Variables: the maximum length of two strings len, length of common subsequence of two strings dp [i] [j]。

```
Begin
    len = maxx(len1 , len2);
    foreach i
          dp[i][0] = 0 ; dp[0][i] = 0 ;
    foreach i
         foreach j
           if(str1[i - 1] == str2[j - 1])
              record public node;
              dp[i][j] = dp[i - 1][ j - 1] + 1 ;
           else
         dp[i][j] = maxx(dp[i - 1][ j ] , dp[i][j - 1]) ;
      return dp[len1][len2];
 End
```

If in a connected undirected graph an arbitrary vertex is deleted, the remaining graph is still connected, then such non-connected graph is called two-pass connected.

Algorithm2 ： DFS algorithm based on social network graph cut algorithm for point

Input：each node m;

Output：cut points in the graph;

Variables: each node corresponds to the dep [m], low [m], flag flag [m], the number of sub-graph can be generated son_num.

```
Begin
      Initialize variables;
      dep[m] = depth;
      low[m] = depth;
      flag[m] = true;
      depth ++;
      foreach i
        if(unvisited)
          DFS(i);
          low[m] = min(low[m],low[i]);
         if(low[i] >= dep[m] && m is not the root node)
                node m is the cut point; record m;
          else if(m is the root)
                son_num ++
          else
                low[m] = min(low[m],dep[i]);
    End
```

Finding a common cut-point algorithm is based on the DFS (Depth-First-Search) algorithm, record the depth dep of each node in the DFS algorithm and the shallowest depth of its descendants can achieve. For each of node: (1) If node m is the root node and has two or more sons, then m is a cut point;(2) If m is not a root and there is a son v, and low[v] >= dep[m], then m is the cut point. The algorithm is as follows.

We can know the importance and the position of each node by finding the cut point in the community. And we can find the importance in the connection of the node in the certain community.

When we analyze the revolution of the community which we have found above, we introduce time properties. Reasonable time period selected has a greater impact on the evaluation of the community revolution; need to consider the reality of the characteristics of the system itself on how to determine the length of the time period. If the divided time period is too long, it is difficult to find the emergence of the network phenomenon. If the divided time period is too short, the number of unstable network nodes and edges may be greatly increased, the results of the community revolution are more susceptible to noise interference, and it will increase computational cost [10-12].

Then, we put forward some parameters to measure the community revolution.

Definition 1. Stability

Stability is the ratio of the static members in the revolution of the community. The higher the stability is, the more of the stable members in the community. If the stability is 1, then the whole community keeps unchanged.

$$R_{stability}(c_i, c_j) = \frac{|c_i \cap c_j|}{|c_i|} \quad (1)$$

Definition 2. Disappearance

Disappearance is the number of the members that disappeared which is in proportion to the members that have disappeared, and inversely proportional to the size of the community itself.

$$R_{disappear}(c_i, c_j) = \frac{|c_i - c_j|}{|c_i|} \quad (2)$$

Definition 3. Growth

Growth is the ratio of the members that increased in the community revolution which is in proportion to the members that are new born, and inversely proportional to the size of the community itself..

$$R_{grow}(c_i, c_j) = \frac{|c_j - c_i|}{|c_i|} \quad (3)$$

Definition 4. Alteration

Alteration reflects the extension of the changes of the members in the community revolution.

$$R_{alteration}(c_i, c_j) = \frac{|c_j \cap c_i|}{|c_i|} \quad (4)$$

Definition 5. Correlation

Correlation shows the relationship between two communities. It is defined as follows:

$$CR(c_i, c_j) = \frac{|c_i \cap c_j|}{|c_i \cup c_j|} \quad (5)$$

If Ci and Cj are the same structure of one community in an adjacent time status, then the relevance between them is called self-correlation. It is obvious that the range of the correlation is [0,1]. If $CR(c_i, c_j) = 1$, the members in the community keep unchanged, if $CR(c_i, c_j) = 0$, the members in the community are totally different.

In this section, the data set are collected from TianYa forum, and the time is from 2010.12. We choose the posts that contain about 10,000 to 20,000 replies, and the time span is long. We use the methods which have been put forward above to find opinion leaders and evaluate the community revolution [13].As the time span of the post is two months, if the time division is one month for only two months. We cannot get the desired results. If it is divided into day time segments, we will have more amount of data analysis; the results are messy and difficult to clearly reflect the results of the evolution of the community. In each network diagram that is extracted, we count the number of nodes, edges, average degree of nodes and other attributes.

Based on level extraction method and the LCS algorithm we get public nodes in the post and we analyze that during the eight weeks there are 81 individuals that are more active speakers; or you can say that 81 nodes in the community structure is relatively stable, they do not change in the stage. These people are in a more central location in the eight weeks.

Later, among 81 individuals, we find that 40 out of 81 nodes can be a cut point with DFS-based method. And we get the connectivity of the certain cut point in the community graph. Therefore, we can say that these 40 individuals found in 81 individuals are more important.

Through this extraction method and the discovery of the opinion leaders approach, we can effectively find the opinion leaders in the post.

TABLE I.

THE CHANGES OF THE PARAMETERS IN MEASURING THE COMMUNITIES

time	stability	dis-appearance	growth	altera-tion	correla-tion
Week1	—	—	—	—	—
Week2	0.24	0.76	0.90	1.66	0.13
Week3	0.37	0.63	0.97	1.70	0.19
Week4	0.28	0.72	0.49	1.21	0.22
Week5	0.01	0.99	0.01	1.00	0.01
Week6	0.14	0.86	0.71	1.57	0.08
Week7	0.08	0.92	0.17	1.19	0.07
Week8	0.33	0.67	1.13	1.80	0.14

Table 1 shows the result of the method proposed in the previous section with certain parameters. From Table 1, we know that the node is changing rapidly, sometimes, the increasing speed high and sometimes low, and disappearance is in a high speed between 0.6-0.9. Therefore, the alteration is also high， resulting in the correlation is between 0.1 to 0.2.

TABLE II.

THE INITIAL INFORMATION OF THE COMMUNITIES AND THE CHANGES OF THE NODES

time	increased number	disappeared number	public node number
Week1	—	—	—
Week2	708	592	188
Week3	867	562	334
Week4	322	869	332
Week5	7	587	7
Week6	10	12	2
Week7	2	2	1
Week8	4	2	1

TABLE III.
THE NUMBER OF THE COMMUNITIES AND THE CLIQUES IN THE POSTS

time	community number	large community number	small community number	Clique number
Week1	16	3	13	1343
Week2	6	2	4	3191
Week3	11	3	8	4756
Week4	10	1	9	1581
Week5	2	1	1	11
Week6	5	1	4	5
Week7	1	0	1	1
Week8	2	0	2	2

Table 2 shows the statistics in single post, including the number of new nodes, the number of disappearance of nodes, and the number of public nodes. The table can be drawn from the post that an increase of 1920 nodes that are new born. However, the number of the disappearance of nodes is more than the number of new born nodes, 2626 nodes are gone. In this basis, the number of participants decreases very fast after four weeks. This explains to some extent, with the time goes by, the post's heat gradually reduces, and the number of people that response reduces.

Table 3 shows the total number of communities, the number of large communities, the number of small communities and the number of cliques in the single post. Here, we define a community is large community if the number of nodes in the community is greater than or equals to 4. Small communities are that the number of nodes is less than 4 communities. From this table we have come to a small number of large communities of 11, while the number of small communities is 42. Therefore, the emergence number of small communities is more than large communities and small communities change faster.

In order to get the number of nodes in the post over time, Figure 2 shows the number of nodes in the distribution of the single post in each week. It can be seen from the figure that the number of nodes in the overall decreasing trend. But in the third week there is a small increase. We get the number distribution of the responses between people (see Figure 3) in a single post. In the third week it reaches the maximum number of responses, then number of responses is in a decreasing trend.

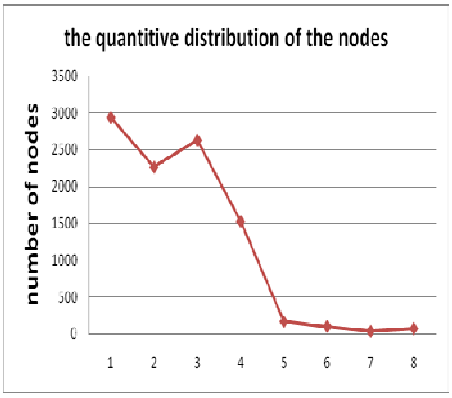

Figure 2. Distribution of nodes in communities

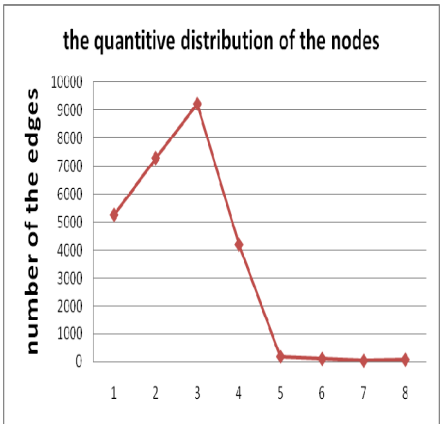

Figure 3. Distribution of edges in communities

V. CONCLUSION

Due to the non-adaptive traditional extraction methods for community social network, we use the level-based community extraction method to extract the opinion leader community, and the data of TianYa Forum are analyzed. The paper analyzes the evolution of the community according to different characteristics of the network by dividing into different time segments and introduces a measuring mechanism and k-groups clustering algorithms in mining opinion leaders and evaluating the revolution of the community.

In the experiment, the level-based community extraction method and the way to find the opinion leaders are effective in extraction of opinion leader community. The association between communities in TianYa Forum is around 0.3 in general. There are a small number of large communities, but they are stable. However, there are a large number of small communities, but they are not stable enough, and easy to be burst and disappear. Above all, the results are more accord with the actual condition.

ACKNOWLEDGMENT

The authors wish to thank Md. Asaduzzaman and Kazuyuki Murase. This work was supported in part by the Shandong Provincial Natural Science Foundation of China (Grant No. ZR2011FL013).

REFERENCES

[1] Changai Hu, Lijun Zhu. The analysis and the evaluation of complicated network software. LNCS, Vol. 13, No. 10, pp.1-5, 2010.

[2] Yunfeng Wang, Hongde Xia, Raomei Yan. The analysis of the social network and the study of the application cases of NetDraw. Modern education technology,18(4):85-89, 2008.

[3] Pothen A, Simon H, Liou K P. Petitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. 11: 430, 1990..

[4] Grivan M, Newman ME J. Community structure in social and biological networks. Proc. Natl. Acad. Sci., 99: 7821-7826, 2001.

[5] NewmanME J, Grivan M. Finding and evaluating community structure in networks. Phys. Rev. E, pp.69-

84,2004.
[6] Toyoda M, Kitsuregawa M. Extracting evolution of web communities from a series of web archives. Proceedings of the fourteenth ACM conference on Hypenext and hypermedia, August 26-30, Nottingham, UK.78-87, 2003.
[7] Gergely Palla,Imre Der´enyi, and Tam´as Vicsek. The Critical Point of *k-groups* Percolation in the Erd´o´s–Re´nyi Graph . Vol. 128, pp.199-211,2009.
[8] Gergely Palla, Albert-L´aszl´o Barab´asi, Tam´as Vicsek. Community dynamics in social networks. Vol.7, No.3 ,pp.273–287, 2007.
[9] Gergely Palla, Albert-László Barabási and Tamás Vicsek, Quantifying social group evolution. pp. 664-667 2007.
[10] Xu Chuanyun, Zhang Yang, Yang Dan, "Ontology based Image Semantics Recognition using Description Logics", IJACT: International Journal of Advancements in Computing Technology, Vol. 3, No. 10, pp. 1-8, 2011.
[11] Chunhua Ju, Jianliang Wei, "Research on Multi-interest Profile Based on Resource Clustering", JCIT: Journal of Convergence Information Technology, Vol. 7, No. 21, pp. 582-590, 2012.
[12] Md Enamul Kabir, Hua Wang. Microdata Protection Method Through Microaggregation: A Systematic Approach. Journal of Software, Vol 7, No 11, 2415-2423, Nov 2012
[13] Mohamed Adel Serhani, Abdelghani Benharref. Enforcing Quality of Service within Web Services Communities Journal of Software, Vol 6, No 4, 554-563, Apr 2011.
[14] Md. Asaduzzaman, Md. Shahjahan, Kazuyuki Murase. Extraction of Interesting Rules from Internet Search Histories.Journal of Software, Vol 6, No 1, 10-19, Jan 2011.

Jianfang Wang is a lecturer at Nanyang Normal University, Henan Province, China. She was born in 1978. She has a B.S and M.S. in computer science. Her recent research interests revolve around information security and computer network.

Xiao Jiao is a lecturer at Nanyang Normal University, Henan Province, China. He was born in 1984. He has a B.S and M.S. in computer science. His recent research interests revolve around information security and computer network.