Chapter 105
Mining Microblog Community Based on Clustering Analysis

Changchun Yang, Hong Ding, Jing Yang and Hengxin Xue

Abstract In order to explore the potential microblog network community structure, this paper proposed an algorithm based on the discovery of community in microblog using clustering analysis. First, according to the characteristics of the microblog network, we defined the network model. Next, we conducted cluster analysis using K-means algorithm based on network model. Finally, LC module degree function was used to determine the optimization community structure. Through experimental comparison, we found that this algorithm takes into account the communication between users. It is conducive to optimizing the conditions of community found and more in line with the actual situation of the microblog network, deriving more accurate result of division of the community.

Keywords Microblog model • Community structure • K-means algorithm • Information group degree • LC module degree

105.1 Introduction

With the arrival of Web2.0, the social network site is gaining popularity. As one of the social network site, microblog has become the focus of many researchers. Now, there are two methods of exploring complex network community structure. One is hierarchical clustering [1–3] in sociology; another is graph segmentation in computer science. Hierarchical clustering is the traditional method of detecting the network community. The GN algorithm was proposed by Girvan and is widely used in hierarchical clustering. The representative method of graph segmentation is Kernighan–Lin algorithm [4] and spectral bisection method [5].

C. Yang (✉) · H. Xue
School of Economics and Management, NJUST, China
e-mail: ycc@cczu.edu.cn

C. Yang · H. Ding · J. Yang
School of Information Science and Engineering, CCZU, China

Z. Zhong (ed.), *Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012*, Lecture Notes in Electrical Engineering 217,
DOI: 10.1007/978-1-4471-4850-0_105, © Springer-Verlag London 2013

K-means algorithm [6] was put forward by MacQueen based on classification of clustering. This algorithm is one of the widely used clustering methods. It is simple, and the algorithm convergence speed is fast. This paper proposed an algorithm of finding microblog network community using K-means algorithm and microblog network attributes. This algorithm defined a concept of group information degree. It dynamically sets the weight of network edge. In line with the minimum relational degree principles to select new clustering center, maximum relational degree principles to pattern classification was used until all the nodes are divided over [7]. The algorithm can be closer to the characteristics of the microblog network and find the cluster center and makes the quality of community greatly improved.

105.2 Problem Definition

105.2.1 Microblog Network Structure

Based on the study of complex networks, the general network structure is divided into non-directed graph structure and unidirectional graph structure. In microblog network, set each user of bloggers for a node. Users have two types of information: follow and follower. Set the follow for the node in-degree and follower for the node out-degree. So, one-way side and two-way side are two types of edges in the network. The phenomenon of "Backed by the face" is very obvious. As can be seen from the relationships between nodes in the network, the microblog network structure is a hybrid digraph.

According to the concept of microblog network community, the existence of the microblog community depends only on the exchange of information between the users. It means that the users switch information with each other between the post and comments, nothing to do with the direction of the follow among users. This paper sets user to node, proposing the concept of node information group degree that means the reciprocal of sum of the number of retransmission and comments between nodes. Node information group of degree can be a very good response in microblog network bloggers and more accurately in the microblog network community mining. On the basis of node information group degree, the microblog network structure is simplified to the non-directed weighted graph.

The microblog network is a triplet $G = (V, E, w)$, where V is a set of nodes, E is a set of edges, and w assigns a weight to each edge. The node information group degree is $_{vi}$, forwarding number in nodes vj and j is r_{ij}, the number of comments between nodes vj and j is c_{ij}, then the expression of vi is

$$d_{ij} = 1/r_{ij} + c_{ij} \tag{105.1}$$

Setting the weight of edge to the node information group degree, the expression of w_{ij} is

$$w_{ij} = d_{ij} \tag{105.2}$$

105.2.2 Node Relation

The node relation [8] between two adjacent nodes in the network of microblog is determined by the weight of their shared edge. The shared edges between adjacent nodes, the smaller, the weight, the greater, the probability of the path is not to transfer information between communities. They have great probability of belonging to the same community. The relations among them are more close, and the node relation is bigger.

As can be seen analysis, node information group degree between the communities is bigger than within the community. Clearly, the smaller the information group degree between nodes i and j, the greater the node relation between them and the their probability of belonging to the same community is higher. Then, we can define the node relation of two adjacent nodes.

$$\text{node Relation}\ (v_i, v_j) = 1 - w_{ij} \tag{105.3}$$

Both adjacent nodes and non-adjacent nodes are the nodes of microblog network. Maybe there are multiple paths or no paths between non-adjacent nodes. Obviously, if the path between two nodes is longer, the node relation is smaller. The node relation between the two non-adjacent nodes is converted to seek the shortest path between two nodes. The shortest path of two non-adjacent nodes is defined as the path of pass edge at least. Namely, the shortest path between nodes is the path containing the least number of nodes among all the paths connecting the two nodes. Therefore, we can use the breadth-first search algorithm to obtain the shortest path between all non-adjacent nodes and then find the maximum node relation between non-adjacent nodes.

Assuming the shortest path of microblog network between non-adjacent nodes vi and vj is $short\ Path\,(vi,\ vj)$, then the node relation between non-adjacent nodes can be expressed as the product of all nodes relation on the shortest path. If the number of shortest path is s between non-adjacent nodes, then choose the largest product as the node relation of non-adjacent nodes.

$$\text{node Relation}\ (v_i, v_j) = \max_s \left\{ \prod_{(v_i, v_k)\ \in \text{shortPath}\ (v_i, v_j)} \text{nodeRelation}\ \ (v_i, v_k) \right\} \tag{105.4}$$

Formulas (105.3) and (105.4) can be used to construct microblog network node relation matrix R.

$$R = [\text{node Relation}\ (vi,\ vj)]_{\ |V| \times |V|} \tag{105.5}$$

Clearly, R is a symmetric matrix, because node relation by itself does not affect the results of the community divided. In order to facilitate the calculation, the value of main diagonal elements is set to the corresponding node degree.

$$R_{|V| \times |V|} = \left\{ \begin{array}{lll} \deg(vi) & i{=}j \quad \text{and} & vi, vj \in V \\ \text{node Relation}\ (vi, vj) & i \neq j \quad \text{and} & vi, vj \in V \end{array} \right\} \tag{105.6}$$

105.2.3 LC Module Degree

The original community module degree (NG module degree) is a metric proposed by Newman to measure the quality of network division. It is not conducive to the measure of the larger situation of the community size differences. However, the size of microblog network community is greatly different, and then it may not achieve the desired results using the NG module degree. In this paper, we use LC module degree; it has the connection with density of the community and cohesion coefficients and does not associate with the sum of the degree of internal nodes. The expression of LC module degree is defined as follows. The value of Q is bigger, and the community structure is more obvious.

$$Q\ (S1,\ S2,\ldots,Sn) = \begin{cases} 0 & n=0 \\ \dfrac{\sum\limits_{i=1}^{n} L(Si)\mathrm{Coh}(Si)}{n} & n>1 \end{cases} \tag{105.7}$$

ni is the number of nodes in community Si, $E(Si)$ is the number of edges in the

$$L(Si) = \begin{cases} 1 & ni{=}1 \\ \frac{2E(Si)}{ni(ni-1)} & ni{>}1 \end{cases} \tag{105.8}$$

community Si.

$A(Si, Sj)$ is the total number of edges in the connection community Si and Sj.

$$\mathrm{Coh}(S_i) = \begin{cases} 0 & E(S_i)=0 \\ \frac{E(S_i)}{E(S_i)+\sum\limits_{j\neq 1} A(S_i,S_j)} & E(S_i)>0 \end{cases} \tag{105.9}$$

105.3 Algorithms

The traditional K-means clustering algorithm uses the principle of maximum distance to select a new cluster centers and use the principle of minimum distance to conduct pattern classification in the mode characteristic vector sets. The principles applied to the discovery of the microblog network community can be understood as the minimum node relation principle to select a new cluster center, and then as the maximum node relation principle to conduct pattern classification until all the nodes are divided. The steps of microblog network community discovery algorithm based on K-means algorithm are as follows.

Input: Adjacency matrix of microblog network

Output: Community structure of microblog network

Step 1. Assume $\text{center} = \emptyset,\ V1 = V0 - \text{center}\ \ k = 2$ According to equation (105.3) and (105.4), the node relation matrix R of microblog network can be calculated.

Step 2. Select the node that has maximum node relation in nodes sets $V1$ as the first cluster center.

$$\text{center} = \text{center} \cup \{cx\}, \; V1 = V1 - \{cx\}$$

Step 3. If $|\text{center}| \neq k$, go to step 4, else go to step 5.

Step 4. Use matrix R to calculate the average node relation between the nodes in sets $V1$ and the nodes in cluster center. Select the node that has minimum node relation as a new cluster center. Go to step 3.

$$r_j = \sum_{i=1}^{|\text{center}|} r_{ji}/|\text{center}|, \quad r_{\min} = \min_j(r_j)$$

$$\text{center} = \text{center} \cup \{v\}, \quad V_1 = V_1 - \{v\}$$

Step 5. $V1 \neq \emptyset$. Compute the node relation between node vj and node in cluster center. Node vj belongs to a cluster that has the maximum node relation. A cluster is a community, output community result of division.

Step 6. Calculate the LC module degree of the current community division. If $Qk \geq Qk - 1$ and then $k = k + 1$, go to step 3, else the algorithm stops.

105.4 Experiments

In order to test the feasibility of the algorithm, this paper used the Sina microblog data from Web site http://www.datatang.com/data/11819. We selected two sets of test cases, each case was built by two groups of deep link user data. The first group selects the star user "YAO Chen" as the original node. Her followings and followers are one deep link, and their followings and followers are two deep link. We collected information from 551 users. The second group selects the ordinary user "Guo Jiu Ye" as the original node. We also collected information from 551 users by using the similar method.

Original K-means algorithm and the algorithm we proposed to test on above test cases were used. Figure 105.1 shows the community division structure of two sets of experimental data by using Ucinet. Figure 105.2 shows the LC module degree trend when network is divided into k communities.

It can be seen from Fig. 105.1, there are 16 communities in Fig. 105.1a, and the community scale is big. Figure 105.1b has 40 communities, but most of them are small group structure. The community structure in Fig. 105.1a is more obvious than in Fig. 105.1b. The original node of the first group data is a leader node. She is more influential. The communication in the two deep link users by her is frequent. Therefore, the community divided structure is obvious. The second group data take an ordinary user as the original node. The two deep link users lack communication, so their community division structure is not very obvious. There are some scattered nodes after the original network is divided into the communities, this is because some users just only follow other users, but have none or little information between them. How to remove these scattered users to get a pure community structure should also be studied in the future research.

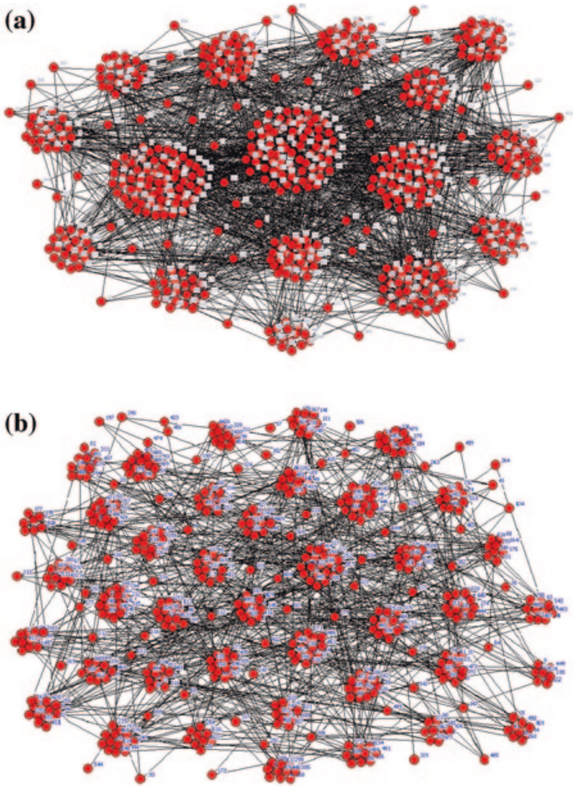

Fig. 105.1 **a** The community division structure of first group data. **b** The community division structure of second group data

It can be seen that the original k-means algorithm takes the cluster structure of all nodes as the final results. The LC module degree is completely a upward trend. The purpose of this algorithm is to obtain an optimal community division structure. When LC module degree reaches the maximum value of community, the division structure is the best result. The original k-means algorithm obtained 12 communities in Fig. 105.2a. Our algorithm obtained 16 communities, and almost all the LC module degrees are higher than in original k-means algorithm. The community structure is more accurate and clear. The line trend of our algorithm is close to the original K-means algorithm, but the number of community is more than the original K-means algorithm. The resulting community structure corresponds to LC module degree greater than the original K-means algorithm. According to the principle, the greater the Q value, the more obviously the community structure shows that the final community structure of our algorithm is more accurate than the original K-means algorithm.

Comparing the combination of the results of two groups of experiments, we take the exchange situation between users as consideration factor of communities division; therefore, we could get a clearer community structure. The more frequent the exchange situation is, the more evident the community structure is. Even when faced

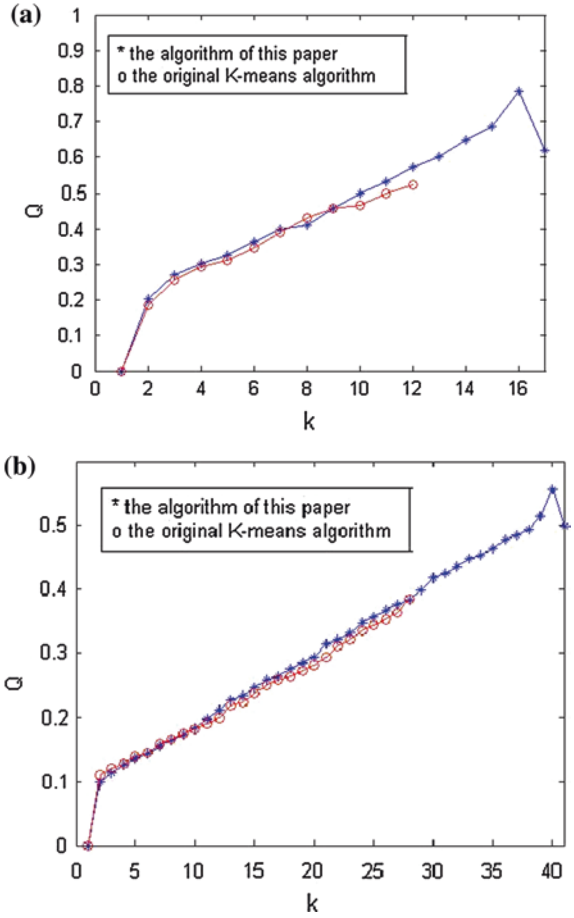

Fig. 105.2 **a** The LC module degree trend of first group data. **b** The LC module degree trend of second group data

with the users lacking communication, the condition of community division has weak advantage; it can be seen from Fig. 105.2b that the number of communities by our algorithm is still more than the original K-means algorithm, and the community structure becomes more clear. The Above analysis shows that the community division structure by our algorithm is better than the original K-means algorithm.

105.5 Conclusions and Future Work

This article quoted the main ideas of the K-means clustering algorithm. It proposed a method to divide microblog network community. The algorithm proposed the concept of node relation. It makes the value more accurate by calculating the network edge weight dynamically. The users' interest similarity degree exists in microblog network. How to put this information into the community discovery algorithm is the research goal in the future.

References

1. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026–033
2. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev E 70(6):066–071
3. Wu JJ, Xiong H, Chen J (2009) Towards understanding hierarchical clustering: A data distribution perspective. Neurocomputing 72:10–14
4. Kim YH, Yoon Y (2009) A new kernighan-lin-type local search for the quadratic assignment problem, vol 22. International conference on scientific computing. CSREA Press, pp 185–189
5. Zhang YP, Wang Y, Zhao S (2010) Detecting communities using spectral bisection method based on normal matrix. Comput Eng Appl 46(27):53–58
6. Ordonez C, Omiecinski E (2004) Efficient disk-based K-means clustering for relational databases. IEEE Trans Knowl Data Eng 16(8):909–913
7. Wang L, Dai GZ, Zhao H (2010) Research on modularity for evaluating community structure. Comput Eng 36(14):75–79
8. Zhao FX, Xie FD (2009) Detecting community in complex networks using K-means cluster algorithm. Appl Res Comput 26(6):88–93