Cluster Comput (2017) 20:1209–1221
DOI 10.1007/s10586-017-0782-y

High-performance social networking: microblog community detection based on efficient interactive characteristic clustering

Ru Wang[1] · **Seungmin Rho**[2] · **Wandong Cai**[1]

Received: 28 September 2016 / Revised: 16 January 2017 / Accepted: 6 February 2017 / Published online: 18 February 2017

Abstract With the development of microblog networks and the popularity of the friend circle, more and more users are linked together to form communities. The microblog community detection is not only the separation of following relationships. The interactive characteristic between users should also be considered. Therefore, in the article the maximum likelihood estimation is used to extract the interactive characteristics clustering. The Link Optimization Comm (LOC) algorithm based on density clustering is proposed to improve the performance the Link Comm (LC) algorithm. By integrating the interactive characteristic clustering into LOC algorithm, the high-performance IC-LOC method to detect potential communities is proposed. Then the complexity in clustering is analyzed. Simulation experiments show that LOC algorithm is better than LC algorithm in time complexity and normalized mutual information evaluation. Compared with LC algorithm in Sina microblog data sets, the IC-LOC method also achieves better performance of community detection. Moreover, the proposed method can effectively detect interactive and potential communities.

Keywords Microblog · Community detection · Interactive characteristic · LOC algorithm

✉ Ru Wang
ruwang@outlook.com

Seungmin Rho
smrho@sungkyul.ac.kr

[1] Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China

[2] Department of Media Software, Sungkyul University, Anyang-si 430-742, Korea

1 Introduction

Microblog is a platform to share, disseminate and acquire information based on user relationships [1]. The user can set up individual communities through WEB, WAP and a variety of clients. The user can update and instantly share a new message in 140 words [2]. Microblog community also refers to a clustering subgroup. The subgroup is a collection of users with similar characteristics. The users exchange the message using the way of "following". Through data mining and analysis in social network, microblog community which is composed of the users with close contact or similar characteristics can be detected [3]. Furthermore, the common characteristic of community members can be excavated. The common characteristic plays an important role in the opinion analysis, product recommendations, and other fields. Because of the mechanism of single direction following in microblog, the user relationship can be abstracted as a directed network (referred to as microblog network). Nodes represent users in the network. The relationship between users is represented as an edge from one user pointing to the follower [4]. In the analysis of complex network, community detection is a very important issue. Because of its great theoretical and practical value, community detection has become a hot issue in network science.

Community detection is also an important issue in the analysis of social network. It helps to sense and understand objects in complex network. Then the further research can be done, such as personalized recommendations [5], friends recommended [6], large-scale network compression solving [7], heterogeneous network analysis [8], the evolution of social networks [9], etc. The double clustering of retweeting and network structure is an important research for precision marketing and personalized recommendation service [10]. In real life, people tend to disseminate the obtained information

from others. Therefore, a good community detection should satisfy the network structure and clustering of interaction parties. The network structure is the bridge between nodes within the community when the information is disseminated. The retweeting is the cause of information dissemination. Thanks to the development of mobile internet, the user scale of microblog and its social influence is growing rapidly. Microblog is a microcosm of the real world. It provides a huge amount of valuable research data for people. People conduct politics [11] and marketing activities [12] using the microblog. Microblog has become a recognized platform to express their views and opinions [13].

In mainstream network, community detection can be categorized in three classes. The first class is based on user content [14]. It extracts characteristics of retweeting in the user content of microblog. Then, users are clustered based on characteristics of interaction. Such a method ignores the effect of microblog network structure, i.e., following relationship, in the information dissemination. The second class is based on user relationship [15]. It extracts friend or following relationship in microblog network. This kind of methods detects the community by network structure. The famous methods include module maximization algorithm [16], GN algorithm [17], Louvain algorithm [18], factions filtering algorithms [19] and link community algorithm [20]. However, these methods do not consider the characteristics of user's interaction. Therefore, the interactive clustering can not be proved. The third class is the integrated approach [21]. It combines the former two kinds of methods. The two communities which have been detected using the former two methods are integrated into one community. The community contains double clustering of interaction and network structure. The method needs to detect community twice, and then integrates two communities. Therefore, the efficiency is too low.

In this paper we use the maximum likelihood estimation to extract the interactive characteristics clustering. And then we propose the Link Optimization Comm (LOC) algorithm based on density clustering to improve the performance the Link Comm (LC) algorithm. The results of experiments show that LOC algorithm is better than LC algorithm in time complexity and normalized mutual information (NMI) evaluation. Afterwards, we propose the IC-LOC method to detect microblog communities. The method integrates the interactive characteristic clustering into LOC algorithm. Then we analyze the complexity and performance of the method. Compared with LC algorithm both interactive clustering and density module, the method can accurately and effectively detect communities.

The paper is organized as follows. Section 2 introduces the related works of community detection in microblog network and expounds the relevant fundamental theory. Section 3 describes the community detection method. Section 4 evaluates the performance of the proposed method and algorithm. The conclusion is given in Sect. 5.

2 Related work

In recent years, with the development of the community in online network, community detection algorithms have been widely studied. Some scholars began to add the attribute information of network nodes into community detection. Steinhaeuser et al. [22] proposed a node attribute similarity (NAS) method. Then he combined it with traditional random walk method. Dang et al. [23] weighted sum of module function and similarity function of node properties. The module was maximized using Louvain algorithm [24]. Community structures had been excavated. Topic model is the most typical method of text clustering algorithm. Latent Dirichlet allocation (LDA) model [25] was proposed based on the probability distribution of multiple topics. Because the user interested in multiple topics with a certain probability distribution. AT (author-topic) model [26] was proposed based on the theme of the probability distribution. It used to detect relationships between users, documents, themes and keywords. Currently more popular community detection algorithm is based on the network structure [27]. According to the relationship between users, the method divided community network into a plurality of sub-communities. The internal structure of sub-communities are closely connected. The relation between the sub-communities is sparse. Kernighan–Lin (KL) algorithm [28] solve the graph partition problem. The algorithm is applied to the community detection in complex network. It is a typical algorithm of graph partition. Using iterative method the graph is decomposed into optimal two subgraphs and repeatedly processed, until a sufficient number of subgraph are obtained. GN algorithm [29] can repeatedly recognize and delete the connection of maximum edge betweenness to cluster the complex network. GN algorithm has high complexity. However, it inspire new ideas of community detection in the complex network. Another GA algorithm has the ability to find the global optimum solution [30]. Therefore, it has good clustering precision. The polymerization method based modularity optimization is a popular algorithm of community detection, and has been expanded to detect weighted network community, directed network community and overlapping community. Although the community detection based network structure (user relationship) can cluster users, it ignores user common interest. Therefore, it can not reflect the interest clustering.

In reality, most of community structure are overlap and hierarchical structure [31]. Interest characteristics of microblog users are diversification. Therefore, community detection of microblog is to detect overlapping community.

CPM algorithm [32] is a popular method to detect overlapping communities. It has been applied in natural and sociology field, and extended to weighting network. However, In CPM algorithm, the community is considered to be a strongly connected cluster. Since the strict definition for community, the effect of community detection is bad in sparse network. In comparison, LINK [17] , link maximum likelihood (LML) [20] and link community (LC) [33] algorithms can detect a better quality of overlapping communities.

In general, $G = (V,E)$ represents social network. $N = \{n_1, n_2, \cdots, n_i\}$ represents the set of nodes. $E = \{e_1, e_2, \cdots, e_j\}$ is the set of edges. i and j denote the number of nodes and edges in network. The arbitrary edge $e_{uv} = (u, v)$ represents two nodes u and v are connected by e_{uv}. In LC algorithm, The edge E is as to a cluster node. According to the edge belonged to the community, LC algorithm clusters the edges, and divides nodes into a plurality of different communities. The several subsets are $S_{LC} = \{LC_i\}$, where $\forall i, j$, $LC_i \cap LC_j = \emptyset$, $\sum_i LC_i = E$. If two edges have a common node, the edges exist similarity. For example, the similarity between $e_{iu} = (i, u)$ and $e_{ju} = (j, u)$ is

$$S(e_{iu}, e_{ju}) = \frac{|\lambda(i) \cap \lambda(j)|}{|\lambda(i) \cup \lambda(j)|} \tag{1}$$

where $\lambda(i)$ is the nodes set which is constituted by i and its neighbor nodes. $\lambda(i) = \{x(i, x) \in E\} \cup \{i\}$.

Assuming a weighted network with N nodes, for any node i LC algorithm has the attribute vector $a_i = (\tilde{A}_{i1}, \cdots, \tilde{A}_{iN})$, and

$$\tilde{A}_{ij} = \frac{1}{k_i} \sum_{p \in n(i)} w_{ip}\sigma_{ij} + w_{ij} \tag{2}$$

where w_{ij} is the weight of $e_{ij} = (i, j)$. $n(i)$ is the set of all neighbor nodes which are connected with node i. k_i is the number of n_i. When $i = j$, $\sigma_{ij} = 1$, otherwise, it is zero. In LC algorithm, the weight w_{ij} of edge e_{ij} represents the relevance of two nodes with a correlation. Typically, the weight is higher, the relevance is greater. Depending on different applications, the implication of w_{ij} is also slightly different. In particular applications, according to the different characteristics of network and different purpose of community detection, w_{ij} can be calculated in different methods. Therefore, Eq. (1) can be written as follows,

$$S(e_{iu}, e_{ju}) = \frac{\mathbf{a}_i \cdot \mathbf{a}_j}{|\mathbf{a}_i|^2 + |\mathbf{a}_j|^2 - \mathbf{a}_i \cdot \mathbf{a}_j} \tag{3}$$

where $\mathbf{a}_i$ and $\mathbf{a}_j$ respectively represents attribute vector of node i and j.

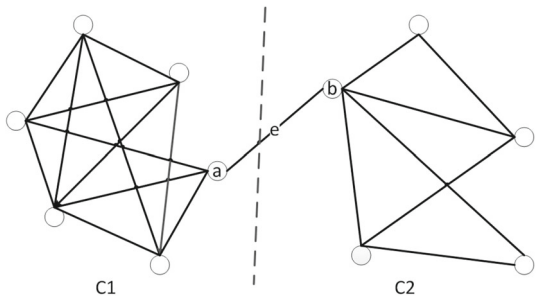

Fig. 1 The example of algorithm unreasonable

LC algorithm clusters the set of edge E using clustering method. Initially, each edges has been respectively considered as a independent link community. Afterwards, two most similar communities continuously are merged, until the following objective function is satisfied.

$$D_c = \frac{2}{M} \sum_c \frac{m_c - (n_c)}{n_c(n_c - 1)/2 - (n_c - 1)} \tag{4}$$

where m_c represents the number of edges in community c. n_c represents the number of nodes in community c. M represents the total number of connected edges in entire network.

However, the result of community detection using LC algorithm is not very accurate. There are two main reasons:

(1) Some communities which have not independent value and large area overlap with other community will appear in results.
(2) In Fig. 1, there is only one edge e connecting between node a and b. If using LC algorithm to divide, two communities will be divided as the dotted line. Node a and b will be divided into two communities. However, this is real division results of microbolg community. In microblog network divided communities depends not only on graph theory, should consider the interaction density between node and edge. If the interaction between node a and b is very frequent, and node a almost don't interact with other nodes. Then node a and b should be classified as a common community. Obviously, LC algorithm do not consider the interactive characteristic of microblog. Only using LC algorithm to detect community is not reasonable.

In order to solve these two problems, LC algorithm is optimized to avoid the problem in *(1)*. Moreover, according to the characteristics of microBlog network, the interactive feature set is added to accurately detect communities.

3 Community detection method of microblog

3.1 LC algorithm optimization

According to following definitions LC optimization algorithm(LOC) can find all the connected communities which satisfy connection and maximum. If an edge does not belong to any connected community, the edge is an isolated edge. Since any two edges are density-connected and symmetrical relationship within the community. Therefore, given the parameters ε and μ, when the community began to detect, the algorithm can star from any edge. The result is uniquely determined.

Definition 1 The neighbor of $e_{uv} = (u, v)$ is $N(e_{uv})$ which is the set of edges respectively connecting node u and v, and not including e_{uv}. That is:

$$N(e_{uv}) = \{\ell = (v, i) \in E \mid i \in N(v)\} \cup \{l = (u, j) \in E \mid j \in N(u)\} - \{e_{uv}\} \quad (5)$$

where $N(v)$ and $N(u)$ respectively represents the neighbor nodes set of node v and u.

Definition 2 Given the threshold ε, $N_\varepsilon(e_{uv})$ is the set of edges of which the similarity are greater or equal to ε in e_{uv} neighbors. That is

$$N_\varepsilon(e_{uv}) = \{\ell \in N(e_{uv}) \mid s(\ell, e_{uv}) \geq \varepsilon\} \quad (6)$$

Definition 3 Given parameter ε and μ, e_{uv} is called the core edge. If e_{uv} in ε field, the number of edges is greater than or equal to μ. That is

$$e_{uv}.Core = ture \Leftrightarrow |N_\varepsilon(e_{uv})| \geq \mu \quad (7)$$

Definition 4 Given parameters ε and μ, when e_{uv} is a core edge and ℓ belongs to ε field of e_{uv}, we are called that directly density reachable of ℓ is to e_{uv}.

$$DirReach(e_{uv}, \ell) \Leftrightarrow e_{uv}.Core = ture \wedge \ell \in N_\varepsilon(e_{uv}) \quad (8)$$

Directly density reachable is an asymmetrical relationship. only when two edges are core edge, symmetrical relationship is satisfied.

Definition 5 Given parameters ε and μ, if there is a edge chain $e_i, e_{i+1}, \cdots, e_j$, making $e_{uv} = e_i$, $\ell = e_j$, and which satisfies directly density reachable of any e_{i+1} to $e_i (i < j)$. It is called the density reachable of ℓ to e_{uv}. That is

$$Reach(e_{uv}, \ell) \Leftrightarrow \exists e_i, e_{i+1}, \cdots, e_j \in E : e_{uv} = e_i \wedge \ell = e_j \wedge \forall i (i < j) : DirReach(e_i, e_{i+1}) \quad (9)$$

The density reachable has transitive, in essence which is a transitive closure of directly density reachable. However, two edges which are not all core edges do not have symmetry.

Definition 6 Given parameters ε and μ, e_{uv} and ℓ are density reachable. If the edge t is existence, the density reachable of e_{uv} and ℓ are all to t. That is

$$Connect(e_{uv}, \ell) \Leftrightarrow \exists t \in E : Reach(t, e_{uv}) \wedge Reach(t, \ell) \quad (10)$$

Definition 7 Given parameters ε and μ, density-connected is symmetric. If the density of edge e_{uv} and ℓ is connection, the density of edge ℓ and e_{uv} also connect. Link community LC is a non-empty subset of edges. The subset satisfies two conditions as follows,

(1) Connection. Any two edges of LC are connected. That is:

$$\forall e_{uv}, \ell \in LC : Connect(e_{uv}, \ell) \quad (11)$$

(2) Maximum. $\forall e_{uv}, \ell \in E : e \in LC \wedge Reach(e_{uv}, \ell) \ell \in LC$

Based on the following definitions, the link community will convert to the corresponding node community set.

Definition 8 Node community C refers to all nodes connected by edges in link community. That is:

$$\forall e_{uv} = (u, v) \in LC \Rightarrow u \in C \wedge v \in C \quad (12)$$

Definition 9 If there are two edges connecting a common node v, and when the edges respectively belong to different link community. The node v is called a overlapping node. That is:

$$Overlap(v) \Leftrightarrow \exists u_1, u_2, LC_1, LC_2 (u_1 \neq u_2 \wedge LC_1 \neq LC_2) : e_{uv} = (u, v) \in LC_1 \wedge e_{uv} = (v, u_2) \in LC_2 \quad (13)$$

3.2 Extraction interactive characteristic set

The real content of a microblog community generally contains three parts: users, relationships,and all kinds of behavior generated by the user. The user behavior includes texts, comments and interactions (retweeting). In these user behavior, only the interaction between users can decide divided community of microblog. Firstly, we make the following definition of microblog network.

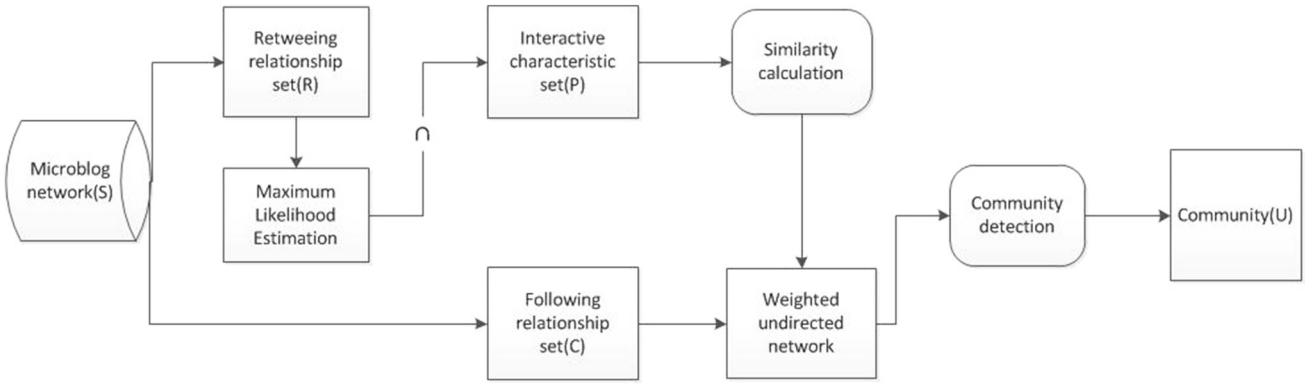

Fig. 2 Framework of community detection in microblog

Definition 10 A microblog generally is represented as $S = (U, L, F)$. Where S represents the microblog community, U represents a set of users. L represents a set of user relationship. F represents a set of interactive characteristic.

To detect microblog community, it is necessary to extract data sets of F. The forwarding behavior of microblog users corresponds a certain probability. The membership strength belonging to a common community affects the probability. Thus, it is necessary to establish a relationship model based forwarding to obtain data sets of F. Using likelihood calculates data sets of F. The provision of interaction model are as follows:

(1) There are a given number of communities and a given number of users.
(2) For each community C and individual x, there is a membership strength parameter F_{xC}. These parameters can be assigned any non-negative. If the value is zero, it means that the individual do not belong to the community.
(3) The probability of edge between u and v in community C is

$$p_C(u, v) = 1 - \exp(-F_{uC} F_{vC}) \tag{14}$$

As usual, the probability of a edge between u and v is that 1 minus the probability which all communities are not prompted edges between them. In other words, the probabilities of the existence edge in each community are independent each other. If there is an edge between two nodes in any community, there is an edge between them. In form, the probability of edge between u and v can be calculated by:

$$p_{uv} = 1 - \prod_C (1 - p_C(u, v)) \tag{15}$$

If in the formula p_C is replaced, there are:

$$p_{uv} = 1 - \exp(-\sum\nolimits_C F_{uC} F_{vC}) \tag{16}$$

Finally, E is the set of edges in the observation graph. As usual, the likelihood of graph will be written in the product of p_{uv} each (u, v) in E. Thus, in the new model, the likelihood formula of the edge set E is:

$$\prod_{(uv)inE} \left(1 - \exp\left(-\sum_C F_{uC} F_{vC}\right)\right) \times \prod_{(uv)notinE} \exp\left(-\sum_C F_{uC} F_{vC}\right) \tag{17}$$

Introducing adjacency matrix A_{uv} between the node u, v, the formula (4) can be written as

$$\begin{aligned} &\prod_{u<v} \frac{\left(\sum_C F_{uC} F_{vC}\right)^{A_{uv}}}{A_{uv}!} \exp\left(-\sum_C F_{uC} F_{vC}\right) \\ &\times \prod_u \frac{\left(\frac{1}{2}\sum_C F_{uC} F_{uC}\right)^{A_{uu}/2}}{(A_{uu}/2)!} \exp\left(-\frac{1}{2}\sum_C F_{uC} F_{uC}\right) \end{aligned} \tag{18}$$

By convention, takes the value $A_{uv} = 1$ if there is an edge between distinct vertices u and v. But $A_{uu} = 2$ for a self-edge, the additional factors of $\frac{1}{2}$ in the second product.

The equation can be simplified by taking the logarithm. Maximization of a function makes logarithm of the function to maximize. Therefore, the formula will takes the natural logarithm. Quadrature becomes summation. In addition, we use $\log(e^x) = x$ in the simplify process. There are

$$\sum_{uv} A_{uv} \log\left(\sum_C F_{uC} F_{vC}\right) - \sum_{uvC} F_{uC} F_{vC} \tag{19}$$

solving such formula maximum value F_{xC} is hard. There is a easier approach as following.

$$\log\left(\sum_C x_C\right) \geq \sum_C q_C \log \frac{x_C}{q_C} \tag{20}$$

where the x_C are any set of positive numbers and the q_C are any probabilities satisfying $\sum_C q_C = 1$. Note that the exact equality can always be get by making the particular choice $q_C = \frac{x_C}{\sum_C x_C}$. Applying expression (20) to (19) gives

$$\log P_{uv} \geq \sum_{uvC} \left[A_{uv} q_{uv}(C) \log \frac{F_{uC} F_{vC}}{q_{uv}(C)} - F_{uC} F_{vC} \right] \quad (21)$$

where the probabilities $q_{uv}(C)$ can be chosen in any way, they satisfy $\sum_C q_{uv}(C) = 1$. Given the true optimal values of F_{uv}, the optimal values of $q_{uv}(C)$ are given by

$$q_{uv}(C) = \frac{F_{uC} F_{vC}}{\sum_C F_{uC} F_{vC}} \quad (22)$$

since there are the values that make out inequality an exact equality. However, given the optimal values of the $q_{uv}(C)$, the optimal F_{uC} can be found by (22), which gives

$$F_{uC} = \frac{\sum_v A_{uv} q_{uv}(C)}{\sum_u F_{uC}} \quad (23)$$

summing this expression over u and rearranging gives us

$$\left(\sum_u F_{uC} \right)^2 = \sum_{uv} A_{uv} q_{uv}(C) \quad (24)$$

and combining with (23) again then gives

$$F_{uC} = \frac{\sum_v A_{uv} q_{uv}(C)}{\sqrt{\sum_{uv} A_{uv} q_{uv}(C)}} \quad (25)$$

Based on maximum likelihood estimation, we get the edge membership strength of microblog community of which the edge is retweeting relation. Thereby, F_{uC} constitutes the set of interactive $F(F_1, F_2, \cdots)$. We will detect the microblog community using interactive characteristic set.

3.3 IC-LOC method

The community detection of microblog is in microblog network to detect the community U of simultaneously clustering with L and F. If F as the research object detect community U, it ignore the important role of L, which can not guarantee that the information within the community can unimpeded spread. In terms of L as a cluster condition detect the community of U, which can not guarantee that the formation community is interactive clustering. Therefore, a reasonable community detection should consider L and F. Existing integrated approach merges two types of community U detected by L and F. Double clustering community U is formed by network structure and interaction. Detecting and merging community twice result in reduced efficiency of the community detection algorithm. The most fundamental reason of detecting community twice do not fully utilize the information and value of L. L as the relationship between users already reflects the existence of U. Therefore, if L is a object of community detection, F is the attribute of L to detect community L. Community L is identified by once community detection, and converted to community U, that will be able to simplify the complexity of community detection.

Since bilateral following (friendship) will better reflect the true social situation. The following relationship discussed in the article refers to bilateral following relationship. The set of user interactive characteristic is a set of weighted values. In order to calculate the interactive characteristic of following relationship, the article defined as follows.

Definition 11 Given a set $A = a_1, a_2, \cdots, a_m$, each element in set all has weight. The weight of the i element a_i is w_{ai}. Then A is the weight set. A also represents as $A = (a_1, w_{a1}), (a_1, w_{a2}), \cdots, (a_m, w_{am})$.

Definition 12 Assumed weight set $A = \{(a_1, w_{a1}), (a_2, w_{a2}), \cdots, (a_m, w_{am})\}$ and $B = \{(b_1, w_{b1}), (b_2, w_{b2}), \cdots, (b_m, w_{bm})\}$, then the intersection of A and B is $A \cap B = \{(e, w_e) | e is the common element of A and B, if e = a_i = b_j, there is w_e = min(w_{ai}, w_{bj})\}$, where the function of $min(\cdot)$ takes the minimum value. For no weight set, it can be assumed that the weight of each element is a fixed constant, such as 1. It can also use the Def.12 to calculate intersection. In microblog network, if using weight set I_i represents the interaction set U_i, the interactive characteristic F_x of following relationship C_x of associated user U_i and U_j can be calculated using the Def.12, namely:

$$F_x = I_i \cap I_j \quad (26)$$

Figure 2 is a basic framework of community detection for microblog network. Firstly, the microblog network *S(U,L,F)* is be divided, and maps *S(C,F)*, where *C(U,L)*. Then, depending on whether there is a common node between C establish the connection relationship between C. Next, according to the interaction characteristic set of F calculate the similarity between interconnected C, and the similarity is set as weight value of connection relationship between C. Weighted undirected network C is established. The problem of community detection is converted to solve the problem of weighted undirected network. The IC-LOC method cluster the interaction, at the same time consider the network structure.

Specifically, the IC-LOC method steps of community detection in micro network as follows:

(1) The set of microblog interaction relationship build a network R. The user is the node, the retweeting relationship of microblog is the edge.

(2) Calculating the interaction characteristic set F. The maximum likelihood estimation of edge in network R be done. Membership strength of each edge will be calculated. Thereby, using the method of Def.11 to intersect obtain the interaction characteristic set $F(F_1, F_2, \cdots)$.
(3) The similarity calculation. For following relationship without a common user, will not calculate the similarity. If there is two following relationship with a common user, the similarity is calculated as follows:

$$Sim(C_1, C_2) = \frac{F_1 \cdot F_2}{|F_1|^2 + |F_2|^2 - F_1 \cdot F_2} \tag{27}$$

(4) The community detection of C. Using LOC algorithm detect community C for above network R.
(5) Forming community. Any C contains two users which have following relationship each other. A community C contains user set of all C. The user set forms U community of corresponding community C. Successively traversing all found community C forms community U.

3.4 Algorithm complexity analysis

Theorem 1 *In a graph which has m nodes and n edges, edges are converted to nodes, nodes are converted to edges. It can form a graph which contains n nodes and* $\frac{1}{2}\sum_{i=1}^{m} L_i^2 - n$ *edges. Where* L_i *is the degree of i node, and there is* $\sum_{i=1}^{m} L_i = 2n$.

Proof Firstly, after that the edges in the original graph convert to nodes, n edges of original graph form n nodes of transformed graph. Then a node in the graph is as the inspection object. If the degree of i node is L_i in the graph, L_i associated edges can form $C_{L_i}^2$ different combinations of two. In transformed graph $C_{L_i}^2$ edges can be formed. Traversing all nodes in the original graph, the number of edges in transformed graph is $\sum_{i=1}^{m} C_{L_i}^2$. Because a edge connects two nodes, in the above traversal process each edge is traversed twice, so there is $\sum_{i=1}^{m} L_i = 2n$. Unfolding $C_{L_i}^2$, there is:

$$\sum_{i=1}^{m} C_{L_i}^2 = \sum_{i=1}^{m} \frac{L_1(L_1 - 1)}{2} = \sum_{i=1}^{m} \frac{L_i^2 - L_i}{2} = \frac{1}{2}\sum_{i=1}^{m} L_i^2 - n \tag{28}$$

Considering a special case, the degree of nodes is 1. Thence, $C_1^2 = \frac{1\times(1-1)}{2} = 0$ bilateral relations between the edges around the node are formed. The theorem is proved. □

Assuming there is microbolg community with m users and n edges of following relationship. After conversing with Theorem 1, the undirected network with n nodes and $\frac{1}{2}\sum_{i=1}^{m} L_i^2 - n$ edges is formed. If the algorithm complexity of the using community detection is $O(f(m, n))$, after transformation, the algorithm complexity is

$$O\left(f\left(n, \frac{1}{2}\sum_{i=1}^{m} L_i^2 - n\right)\right) \tag{29}$$

If using LOC algorithm, we should first analyze the time complexity of the algorithm. Supposing G network contains n nodes and m edges. The first step of LOC algorithm is to traverse each edge and find the edge set of directly density reachable. It only need check neighbor set of each edge. Its time complexity is $O(\sum_i d(e_i))$. Where $d(e_i)$ is the number of e_i edge neighbor. Supposing $e_i = (u_i, v_i)$, then $d(E_i) = k(u_i) + k_{v_i}$, $k(u_i)$ and $k(v_i)$ represent the degree of node u_i and v_i. Thus, assuming the average degree of nodes in the network is k. The time complexity of the first step is $O(km)$. The second step processes each edge in the network. Connected two nodes are assigned to the appropriate node community. Its time complexity is $O(m)$. In lots of real networks $k \ll m$, m and n are linear. Therefore, The overall time complexity of LOC algorithm is $O(m)$ or $O(n)$. After conversion by Theorem 1, the algorithm complexity is $O(f(n, \frac{1}{2}\sum_{i=1}^{m} L_i^2 - n)) = O(n)$. Correspondingly, in Sina microblog spare network, because of $m \sim n$, its time complexity is $O(f(n, \frac{1}{2}\sum_{i=1}^{m} L_i^2 - n)) = O(n) = O(m)$.

4 Experiment and analysis

4.1 Experimental data

In this paper, according the API of Sina microblog, we crawl the data of Sina microblog. The specific method is to crawl all users in following list of one microblog user, and then to crawl user relationship list and label information. After removing the border points, we get each user information including following list and retweeting relationship. Since the retweeting relationship of users may change with time, we only crawl latest 100 retweeting messages in each user. Therefore, this article randomly selects three different users whose the number of friends is $50 \sim 100$ from microblog network as a seed user. We respectively crawl the user ID, the following and the retweeting relationship to form three data sets. In data set $S1$, there are about more than 600 users and 2000 user relationships. In data set $S2$, there are about more than 800 users and 8000 user relationships. In data set $S3$, there are about more than 200 users and 1500 user relationships. From Figs. 3, 4 and 5 the network structure of crawling three data sets. They respectively represent the network basic structure of $S1$, $S2$ and $S3$. The figure shows that the nodes of the network formed by three data sets are

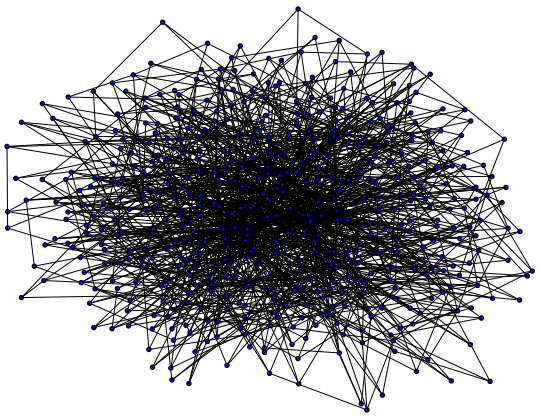

Fig. 3 Microblog network structure of S1

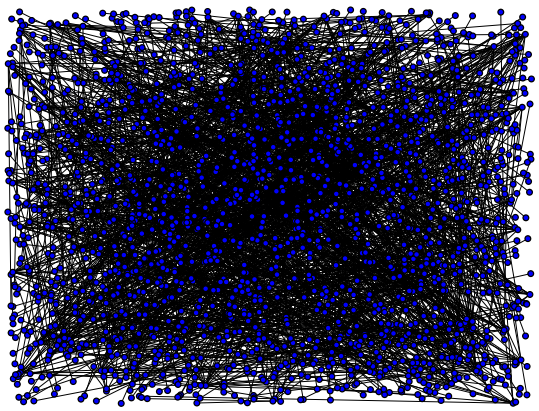

Fig. 4 Microblog network structure of S2

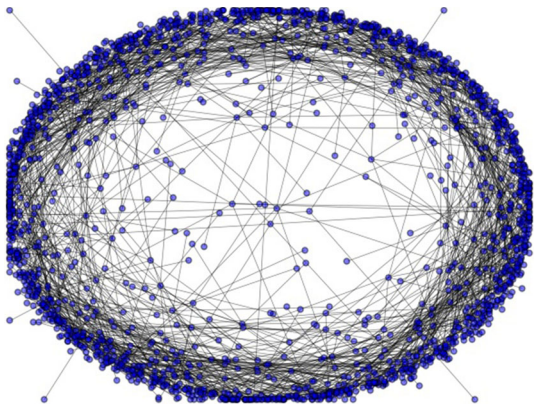

Fig. 5 Microblog network structure of S3

communicated with each other. There are not isolated nodes and groups of nodes.

Figures 6, 7 and 8 show the degree distribution of the three data sets. From the figures we can see that the degree of experimental data in three groups are power law distribution. Therefore, the microblog networks which are formed by the three experimental data sets are scale-free networks [34].

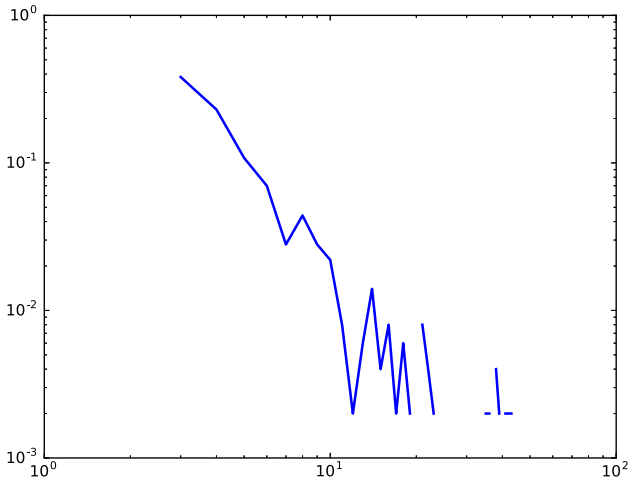

Fig. 6 The degree distribution of S1

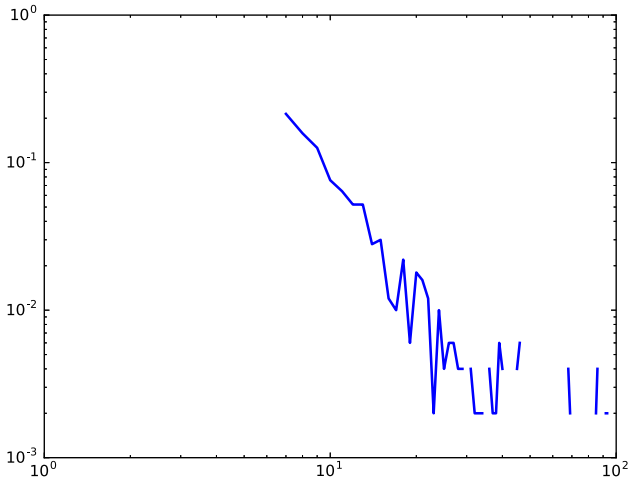

Fig. 7 The degree distribution of S2

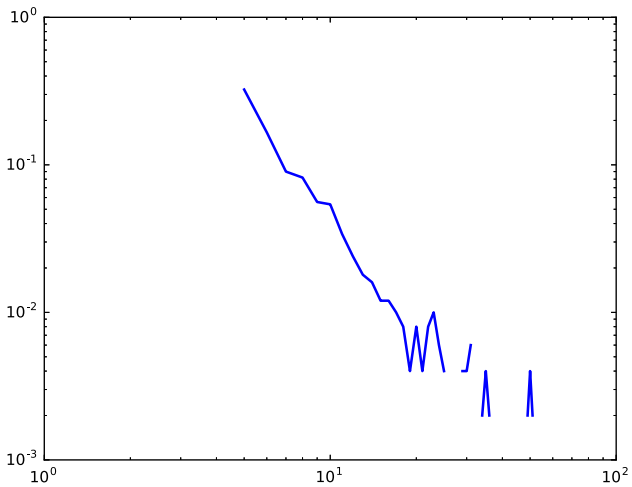

Fig. 8 The degree distribution of S3

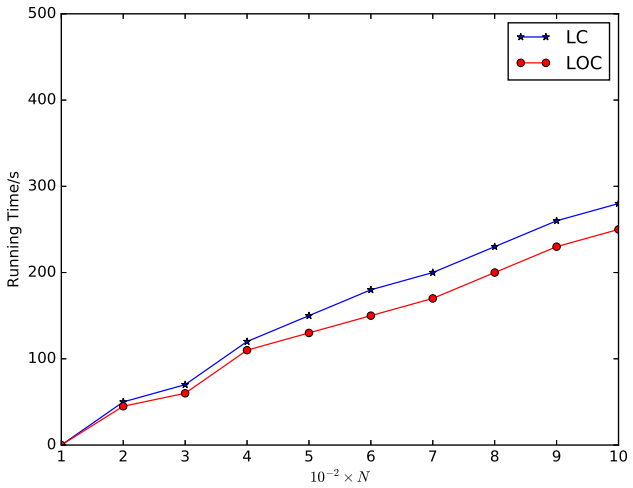

Fig. 9 Time efficiency comparison of two algorithms

Table 1 Information of the standard network

Name	N	k	Minc	Maxc	On
L1	1000	10	50	50	100
L2	1000	10	50	50	500
L3	1000	20	100	100	100
L4	1000	20	100	100	500

4.2 Performance analysis of LOC algorithm

Since the LC algorithm calculates the similarity between edges. In order to better control the experiment, this experiment calculates the the interaction similarity between nodes and interaction sets of following relationship. In this article we improved the LC algorithm to propose the LOC algorithm. In order to better compare the performance of two algorithms, we use the normalized mutual information (NMI) evaluation criteria which is given in literature [35].

First of all, we compare the time efficiency of two algorithms. We give 10 standard networks. The value of nodes sequentially are 1000, 10,000, and other parameters are the same. The operation time of two algorithms in 10 standard networks is shown in Fig. 9. In the figure, the abscissa represents the number of nodes, the ordinate represents the running time of algorithms. As can be seen from Fig. 9, the time efficiency of LOC is higher than LC algorithm. When $N < 5000$, the time efficiency of two algorithms are almost same. However, when $N > 5000$, the time efficiency of LOC is better than LC algorithm.

Next we evaluate the detection result of two algorithms. We respectively generate four groups of the standard network. The information of the standard network is shown in Table 1

where N represents the number of nodes. k represents the average degree of nodes. $minc$ represents the number of nodes in small communities. $maxc$ represents the number of nodes in the largest community. on represents the number of nodes. om represents the number of community which overlapping nodes link. Each algorithm selects the maximum from the value of NMI in different parameters as the final result. Figure 10 shows the detection results of two algorithms respectively in four groups of the standard network. The abscissa represents the value of om. The ordinate represents the quality of community detection NMI. As can be seen from the figure, comparing the NMI value of LC, the NMI of LOC has a very obvious superiority. The LOC algorithm more reasonably and accurately process isolated edges in network. The valid link community is converted to the corresponding node community.

4.3 Interaction characteristic analysis

Interaction clustering refers to the characteristics which the user in community frequently retweets messages to interact with other users. In order to better describe the interaction characteristics, we define as follows:

Definition 13 Interaction clustering index is used to describe the ratio of interaction similarity between users and the entire community similarity. If there are any two users U_i and U_j in microblog network, the interactive similarity is $\phi(i, j)$. The interactive clustering E is:

$$E = \frac{\sum_{in\ one\ community\ node\ i,j} \phi(i, j)}{\sum_{all\ nodei,j} \phi(i, j)} \tag{30}$$

Obviously, the value of E is higher, the detected communities have better interaction clustering, the community detection algorithm is more superior. Although E describes the overall interaction clustering, it can not describe the single interaction clustering. Therefore, we define interaction average index to describe the single interaction clustering.

Definition 14 Interaction average index describes the average value of community total similarity. If the community C has $|C|$ users, the interaction average index e is:

$$e = \frac{\sum_{i,j \in C} \phi(i, j)}{|C|} \tag{31}$$

e is higher, the interaction clustering is better.

Figure 11 shows the interaction average index of two algorithms in three sets. The figure visually depicts the comparison of interaction average index in ten communities using two algorithms. Obviously, the method proposed in this article is better than LC algorithm in the interaction average index of a single community. Therefore, the method proposed in this article has a better interaction clustering in detected community.

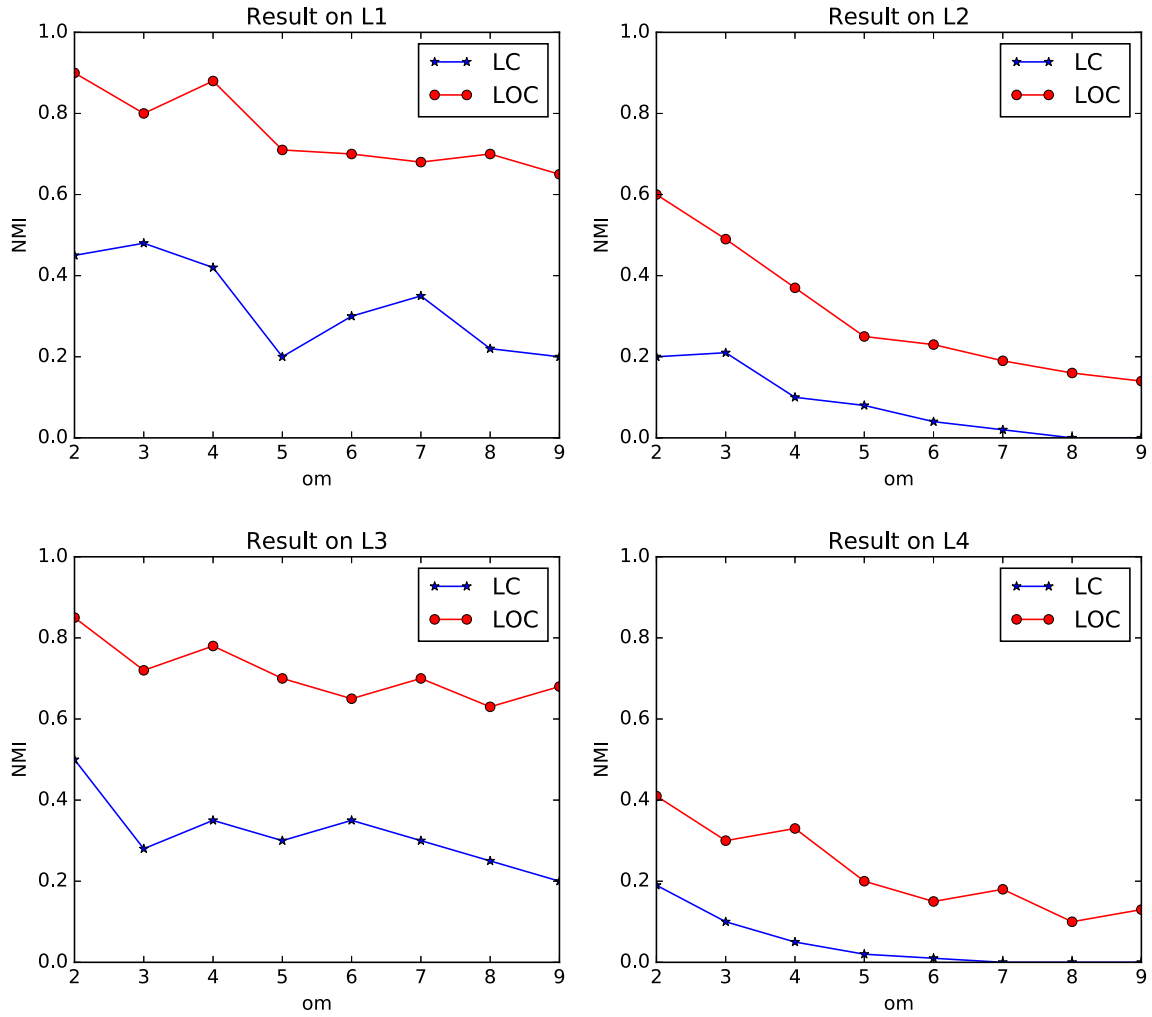

Fig. 10 The detection results of two algorithms respectively in four groups of the standard network

4.4 Results comparison

Community density is the standard to partition community level for LC algorithm. However, it is not a good evaluation. Therefore, in this article the community density could not be an evaluation. Modularity is a commonly evaluation to evaluate aggregation of nodes. Suppose the membership strength of node i and node j for community c respectively is ρ_{ic} and ρ_{jc}, we define the function $F(\rho_{ic}, \rho_{jc})$:

$$F(\rho_{ic}, \rho_{jc}) = \frac{1}{(1 + exp(-f(\rho_{i,c})))(1 + exp(-f(\rho_{j,c})))} \tag{32}$$

where $f(x) = 60x - 30$. According the function F, we calculate Q as follows:

$$\xi_{l(i,j),c} = F(\rho_{i,c}, \rho_{j,c}) \tag{33}$$

$$\xi^{out}_{l(i,j),c} = \frac{\sum_{j \in V} F(\rho_{i,c}, \rho_{j,c})}{|V|} \tag{34}$$

$$\xi^{in}_{l(i,j),c} = \frac{\sum_{j \in V} F(\rho_{i,c}, \rho_{j,c})}{|V|} \tag{35}$$

$$Q = \frac{1}{m} \sum_{c} \sum_{i,j \in V} \left[\xi_{l(i,j),c} A_{ij} - \xi^{out}_{l(i,j),c} \xi^{in}_{l(i,j),c} \frac{k^{out}_i k^{in}_j}{m} \right] \tag{36}$$

where A is the adjacency matrix of network to describe the connection between node i and node j. If node i and node j have connection, $A_{ij} = 1$, else $A_{ij} = 0$. Obviously, when the number of internal edges in divided community is greater, the community structure is more obvious.

Figure 12 shows the trend of Q based on ε using the method which is proposed in three sets. ε is proposed in Def.2. ε is a very import parameter to evaluate LOC algorithm. IC-LOC method is based on LOC. As can be seen from

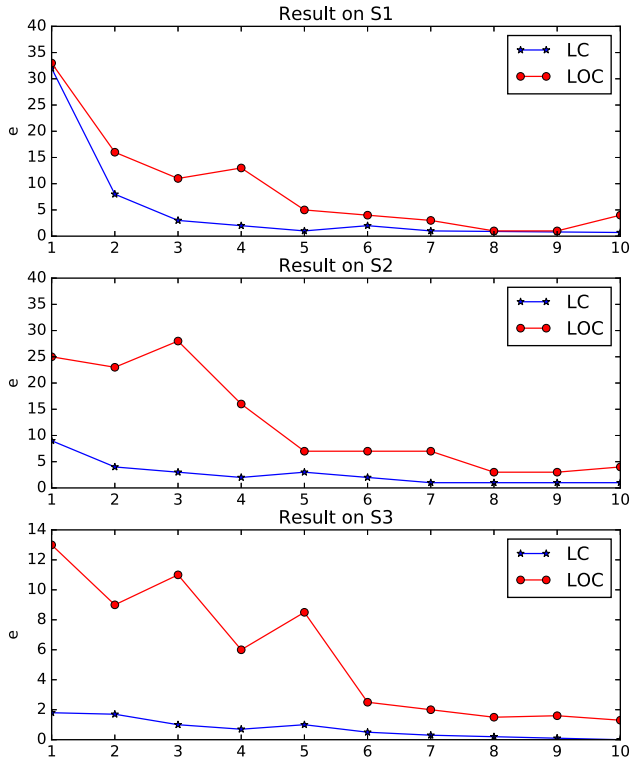

Fig. 11 The interaction average value of e

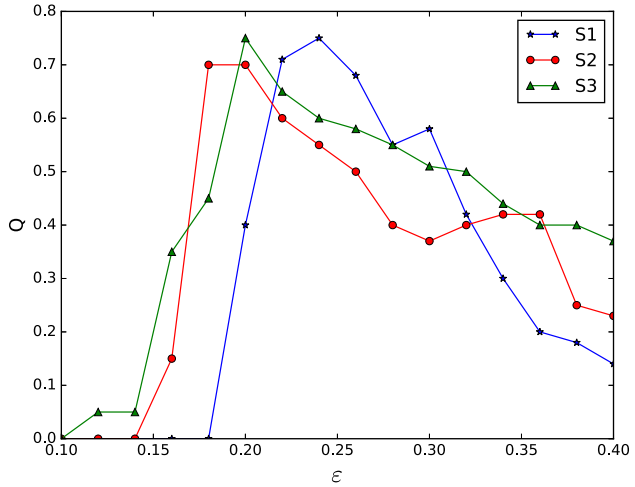

Fig. 12 The Q value with different ε

the figure, the IC-LOC method arises the extreme value of Q under certain ε in a real network, and it appears Q value is decremented on each side of ε. Therefore, in practice, we need set up a initial parameter ε_0. According to the change of Q, we use climbing strategy to adjust ε until it reaches the optimal Q. The proposed method of community detection is better than LC algorithm in the overall network structure.

Finally, we apply the proposed method to a real network. We extract a part of S_1. Figure 13 shows the community structure using the proposed method. In the figure the red

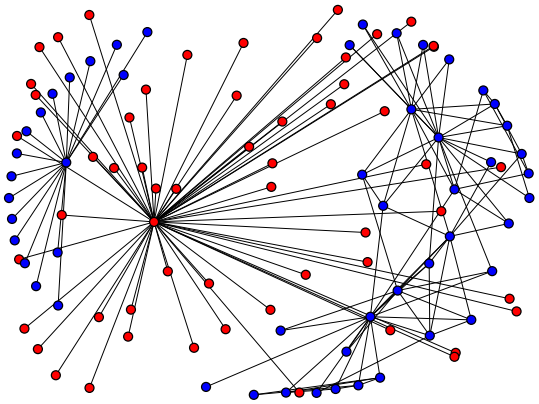

Fig. 13 The result of community detection

node is the network community using the proposed method. The red label community has been able to detect, because they interact frequently between the nodes in a short time. Although the connection of red nodes is not very close, they interact frequently. However, the red nodes are assigned to the same community. The interactive characteristic which is added into community detection is different with the traditional community detection. Its structure may not close, but the similarity is high. Therefor, in real networks it is more authenticity and accurate.

5 Conclusion

Based on the analysis of existing community detection, LC algorithm are shortcomings. Therefore, the community detection based on interactive characteristic of sina microblog is proposed. And LOC algorithm which improves the LC algorithm is proposed. The time complexity of LOC algorithm is analyzed. Using MLE the retweeting behavior is calculated and the interactive characteristic set of users is extracted. On the basis of LOC algorithm adding interactive characteristic set, a new method of community detection (IC-LOC)is proposed. In three groups of microblog real data sets, this article analyzes interaction clustering and similarity module of the proposed method by comparing LC algorithm. Finally according to detect communities in real networks, the proposed method conforms the real situation in microblog network, and detects the community more comprehensive. The IC-LOC method is applied to detect Sina microblog community, and detecting community structure is more optimum. This method for other online communities is applicable, such as Twitter, Google+ and Tencent micoblog,etc. In subsequent work, we will test and verify it in more data sets and detect better user communities.

References

1. Huang, L., Wang, S., Hsu, C.H., Zhang, J., Yang, F.: Using reputation measurement to defend mobile social networks against malicious feedback ratings. J. Supercomput. **71**(6), 2190–2203 (2015)
2. Wang, S., Huang, L., Hsu, C.H., Yang, F.: Collaboration reputation for trustworthy web service selection in social networks. J. Comput. Syst. Sci. **82**(1), 130–143 (2016)
3. Thompson, P.: The digital natives as learners: technology use patterns and approaches to learning. Comput. Educ. **65**, 12–33 (2013)
4. Wang, R., Rho, S., Chen, B.W., Cai, W.: Modeling of large-scale social network services based on mechanisms of information diffusion: Sina weibo as a case study. Future Generation Computer Systems (2016)
5. Lim, K.H., Datta, A.: Following the follower: detecting communities with common interests on twitter. In Proceedings of the 23rd ACM Conference on Hypertext and Social Media, pp. 317–318. ACM, Milwaukee (2012)
6. Chen, B.-W., Wang, J.-C., Wang, J.-F.: A novel video summarization based on mining the story-structure and semantic relations among concept entities. IEEE Trans. Multimedia **11**(2), 295–312 (2009)
7. Bhattacharya, S., Henzinger, M., Nanongkai, D., Tsourakakis, C.: Space-and time-efficient algorithm for maintaining dense subgraphs on one-pass dynamic streams. In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, pp. 173–182. ACM, Milwaukee (2015)
8. Amelio, A., Pizzuti, C.: A cooperative evolutionary approach to learn communities in multilayer networks. In International Conference on Parallel Problem Solving from Nature, pp. 222–232. Springer, Cham (2014)
9. Fagnan, J., Rabbany, R., Takaffoli, M., Verbeek, E., Zaïane, O.R.: Community dynamics: event and role analysis in social network analysis. In International Conference on Advanced Data Mining and Applications, pp. 85–97. Springer, Cham (2014)
10. Chen, B.-W., Ji, W.: Intelligent marketing in smart cities: crowdsourced data for geo-conquesting. IEEE IT Prof. **18**(4), 18–24 (2016)
11. Larsson, A.O., Moe, H.: Studying political microblogging: Twitter users in the 2010 swedish election campaign. New Media Soc. **14**(5), 729–747 (2012)
12. Lim, K.H., Datta, A.: Finding twitter communities with common interests using following links of celebrities. In Proceedings of the 3rd International Workshop on Modeling Social Media, pp. 25–32. ACM, Milwaukee (2012)
13. Kim, E., Sung, Y., Kang, H.: Brand followers retweeting behavior on twitter: how brand relationships influence brand electronic word-of-mouth. Comput. Hum. Behav. **37**, 18–25 (2014)
14. Bao, J., Zheng, Y., Wilkie, D., Mokbel, M.: Recommendations in location-based social networks: a survey. Geoinformatica **19**(3), 525–565 (2015)
15. Newmann, M.E.J.: Communities, modules and large-scale structure in networks. Nat. Phys. **8**(1), 25–31 (2012)
16. Newan, M.E.J.: Spectral methods for community detection and graph partitioning. Phys. Rev. E **88**(4), 042822 (2013)
17. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. Knowl. Inf. Syst. **42**(1), 181–213 (2015)
18. Le Martelot, E., Hankin, C.: Fast multi-scale detection of relevant communities in large-scale networks. Comput. J. bxt002 (2013)
19. Li, J., Wang, X., Cui, Y.: Uncovering the overlapping community structure of complex networks by maximal cliques. Physica A: Stat. Mech. Appl. **415**, 398–406 (2014)
20. Gopalan, P.K., Blei, D.M.: Efficient discovery of overlapping communities in massive networks. Proc. Natl. Acad. Sci. **110**(36), 14534–14539 (2013)
21. Zhang, Z., Li, Q., Zeng, D., Gao, H.: User community discovery from multi-relational networks. Decis. Support Syst. **54**(2), 870–879 (2013)
22. Chen, B.-W., Chen, C.-Y., Wang, J.-F.: Smart homecare surveillance system: behavior identification based on state transition support vector machines and sound directivity pattern analysis. IEEE T. Syst. Man Cy.: Syst. **43**(6), 1279–1289 (2013)
23. Chen, B.-W., Tsai, A.-C., Wang, J.-F.: Structuralized context-aware content and scalable resolution support for wireless VoD services. IEEE T. Consum. Electr. **55**(2), 713–720 (2009)
24. Takemura, S., Bharioke, A., Lu, Z., Nern, A., Vitaladevuni, S., Rivlin, P.K., Katz, W.T., Olbris, D.J., Plaza, S.M., Winston, P.: A visual motion detection circuit suggested by drosophila connectomics. Nature **500**(7461), 175–181 (2013)
25. Anandkumar, A., Liu, Y.k., Hsu, D.J., Foster, D.P., Kakade, S.M.: A spectral algorithm for latent dirichlet allocation. In Advances in neural information processing systems , pp. 917–925 (2012)
26. Blei, D.M.: Probabilistic topic models. Commun. ACM. **55**(4), 77–84 (2012)
27. Wang, R., Cai, W., Shen, B.: The study of the dynamic model on KAD network information spreading. Telecommun. Syst. 1–9 (2015)
28. Kappes, J., Andres, B., Hamprecht, F., Schnorr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Lellmann, J., Komodakis, N., et al.: A comparative study of modern inference techniques for discrete energy minimization problems. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1328–1335 (2013)
29. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state-of-the-art and comparative study. ACM Comput. Surv. (csur) **45**(4), 43 (2013)
30. Wang, R., Cai, W.: A sequential game-theoretic study of the retweeting behavior in sina weibo. J. Supercomput. **71**(9), 3301–3319 (2015)
31. Ahn, Y.-Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
32. Fan, J., Chen, X.: General clique percolation in random networks. EPL (Europhysics Letters) **107**(2), 28005 (2014)
33. Kim, Y., Jeong, H.: Map equation for link communities. Phys. Rev. E **84**(2), 026110 (2011)
34. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)
35. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New J. Phys. **11**(3), 033015 (2009)

Ru Wang received the B.S. degree in 2008 from Xidian University (Xian, Shaanxi China), and the M.S. degree in 2011 from the Changan University (Xian, Shaanxi China). She is working towards the Ph.D. in Northwestern Polytechnical University (Xian, China) of Computer Science and Technology from September 2011. Her research focuses on P2P, Social Network and Complex Network.

Seungmin Rho was with Carnegie Mellon University, USA. He is currently a Faculty Member with the Department of Media Software, Sungkyul University, South Korea. His research interests include database, big data analysis, music retrieval, multimedia systems, machine learning, knowledge management, and computational intelligence.

Wandong Cai received the Ph.D. in computer science and technology from Northwestern Polytechnical University, Xi'an, China. He is currently a Professor with the School of Computer Science and Technology, Northwestern Polytechnical University. He is the author of more than 200 papers in his areas of interest, including the measurement and analysis of complex networks, security monitoring of network information, security evaluation of information system. He is currently a Director of the Network Security Institute and a Committee Member of the Computer Security Professional Committee of China.