A novel overlapping community mining algorithm for micro-blog platform

Zhang Zhaoyin[*]

School of Computer Science and Technology in Heilongjiang University 150010, China

Received 1 November 2014, www.cmnt.lv

Abstract

This paper concentrates on the problem of overlapping community mining for micro-blog platform, which is an important problem in social network mining. Firstly, we convert of overlapping community mining to a weighted graph computation problem, in which nodes represent users and vertex denotes the relationship between users. Secondly, we introduce the concept of user influence to solve the problem of overlapping community mining, which is a main innovation point in this paper. To calculate the user influence in the micro-blog platform, two types of micro-blog information are utilized (that is, user properties and micro-blog properties), and then the analytic hierarchy process is used to calculate the weight of each influencing factor. Furthermore, user properties contain user ID, user type, attention number, number of fans, number of micro-blog, number of mentions and so on. On the other hand, micro-blog attributes contain micro-blog number, publishing date and time, forwarding number, comment number and so on. Thirdly, a weighted network based overlapping community mining algorithm is proposed, in which the original overlapping communities are discovered in advance, and then final results are obtained by expanding the original ones. Finally, to testify the effectiveness of the proposed, experiments are conducted on several datasets and compared with other related works. Experimental results demonstrate that the proposed algorithm can detect overlapping communities in micro-blog platform with high accuracy, and our algorithm is suitable to be modified to run in the parallel mode, hence, the large-scale overlapping communities can also be solved by this proposed algorithm.

Keywords: Overlapping community, Micro-blog platform, weighted graph, User influence, Membership degree.

1 Introduction

With the rapid development of the Internet, micro-blog platform services such as Facebook and Twitter have become more and more popular. Micro-blog platform can provide an effective way for users to share multimedia contents and communicate with others. Particularly, many popular social networking websites such as Facebook and MySpace can provide support micro-blogging services [1]. The main differences between micro-blog platform and traditional blog websites lie in that the limited message size. Although micro-blog platform is constructed according to the requirements of human communication, they have many features that provide opportunities for data mining [2][3]. Hence, in recent years, there are many researches related to micro-blog data mining, such as event discovery, user influence evaluating, information dissemination trend, and community mining.

Community is regarded as an important attribute of the real social networks, because community greatly influence the whole complex system. Despite there are many difficulties in community mining, a lot of technologies have been presented, such as random walks, spectral clustering, modularity maximization, differential equations, and so on[4]. Detecting communities in micro-blog platform is an important task for understanding the internal structures of complex networks. In recent years, several methods have been proposed to detect disjoint communities. The complex systems can be modelled as graphs or networks. Community mining refers to discover community structures. Particularly, if nodes importance or relationships between nodes are not equal, the weighted graph should be utilized [5][6]. In recent years, the community structure mining in complex networks has attracted many attentions of researchers. However, in real complex networks, graphs nodes are often shared by at least two communities, that is, overlapping community. Detecting overlapping community structures is a key step to study the internal structure of complex networks.

In the existing studies of community mining, many attentions have been concentrated on identifying disjoint communities. The disjoint community mining problem supposes that the network can be separated into dense parts, in which graph nodes have more connections than the ones out of this region. However, in fact, users in micro-blog platforms are characterized by many attributes. For example, a micro-blog user may be subordinate to several social communities, such as family, friends or workmates. Furthermore, in the social networks, a user can join to any communities if he wants. Hence, it can be concluded that overlapping is an important feature in real-world social networks [7]. Based on the above analysis, there is growing interest in overlapping community mining research which can find several clusters that are not necessarily disjoint [8-9]. In this problem, there are several nodes belonged to more than one community.

[*] ***Corresponding author's*** e-mail: zhaoyingaga@126.com

The main innovations of this paper lie in that we introduce the concept of user influence in the community mining problem, and the relationships between users in the micro-blog platform can be estimated more accurately. The rest of the paper is organized as follows. Section 2 illustrates related works of this paper. In section 3, structure of the overlapping communities and some important definitions are proposed. As an important step in overlapping community mining, the method of user influence calculating is proposed in section 4. Section 5 illustrates the proposed overlapping community mining algorithm based on weighted network. To demonstrate the performance of our algorithm, in section 6, experiments are conducted on several datasets, and experimental results testify the effectiveness of our algorithm. Finally, the conclusions are drawn in section 7.

2 Related works

In this section, we illustrate related works of overlapping community mining on micro-blog platform by two aspects, and analyze the difference between former works and this paper. In the first aspect, we propose some typical works that analyze the characteristics of communities in micro-blog platform. Twitter refers to a popular online social network and micro-blog platform that enables users to send and receive short 140-character text messages, named "tweets". Twitter was developed in March 2006 by Jack Dorsey in July 2006. Currently, Twitter is the most popular micro-blog platforms. Users can communicate with others through this system using its interface. Users can put forward messages in public or available only to friends. In the following years, twitter developed rapidly gained more and more attentions. Detecting and mining community structures in twitter are of great importance to understand the internal structure of this micro-blog platform, and related woks are given as follows.

> Adebayo et al. studied on the Structure of the BBC News-Sharing Community on Twitter. The authors found that 1) a majority of BBC audiences use English, Spanish, Russian, and Arabic to receive news, and 2) Twitter users mainly follow the ones sharing news in the same language, and these users are usually located in geographical regions in which the specific language is native[10].

Based on the Working Group 1 report published on the Intergovernmental Panel about "Climate Change", Pearce et al. concentrated on how Twitter users formed communities around their conversational connections. Furthermore, the authors find that, in this forum, users are most likely to converse with users having similar views. However, qualitative analysis proved that the emergence of a community of Twitter users, predominantly based in the UK, where greater interaction between contrasting views took place [11].

Herdagdelen et al. analyzed the characteristics of Geography and Politics of News-Sharing Communities in Twitter. In this paper, the authors map the social, political, and geographical properties of news-sharing communities on Twitter. Furthermore, they track user-supplied messages which contain links to New York Times online articles and then they label users with the topic of links. If users are clustered based on who follows whom in Twitter, they discover social groups separate by if they are interested in local, national or global problem[12].

Ikeda et al. proposed a novel demographic estimation algorithm for profiling Twitter users, using tweets and community relationships. Furthermore, a hybrid text-based and community-based approach for the demographic estimation of Twitter users is given as well. In this algorithm, demographics are obtained by searching the tweet history and then classifying followers or followees [13].

Kim et al. utilized clustering technologies in twitter community detection and issue extraction. Particularly, they cluster nodes in Twitter to find a set of user with similar interest, named community, and they also utilized the Louvain algorithm and proposed a partitioned Louvain algorithm as its modified version[14].

For the micro-blog platform, as interests of users are diverse, they may attend more than one group. Therefore, to study on the internal structure of the complex network in micro-blog platform, mining and detecting the overlapping communities is of great importance. In the following parts, some typical works about mining and detecting the overlapping communities in micro-blog platform are listed as follows.

Rhouma et al propose a method named DOCNet to detect overlapping communities in complex networks. The main innovation of this algorithm is to obtain an initial core and add suitable nodes to expand it until some conditions are satisfied [15].

Chen et al. present a sighed probabilistic mixture to detect overlapping community in signed networks. Different from the existing works, the advantages of the proposed algorithm lie in that this method can provide a soft-partition solution for signed networks, and can calculate soft memberships for nodes [16].

Cui et al. defined three test conditions for overlapping nodes and then illustrate a fast overlapping community detection algorithm which can correct errors by itself. Furthermore, the authors improve the bridgeless function that can evaluate the extent of overlapping nodes [17].

In paper [18], local random walk and a new distance metric are defined. Next, the dissimilarity index between each node of a network is computed in advance. To keep the original node distance, the network structure is mapped to a low-dimensional space through the multidimensional scaling algorithm. Then, the fuzzy c-means clustering algorithm is utilized to detect fuzzy communities in the complex network [18].

Wang et al. use the generative model to detect overlapping communities. In this method, the community memberships of each node are computed by a probabilistic approach based on several parameters. Particularly, the node participation degrees in each community can be calculated effectively [19].

Badie et al. presented a new algorithm to detect both overlapping and non-overlapping community structures in complex systems. The proposed algorithm utilizes several agents for investigation of the input network, and agents can implement the investigation process with nodes closeness[20].

Shi presented a new approach to detect overlapping communities. Innovations of this paper lie in that link clustering is used in overlapping community detection. The proposed algorithm exploits genetic operation to cluster edges of the graph. Moreover, this algorithm can automatically obtain the optimal community number, which is an important factor in community detection[21].

Zhang et al. proposed a symmetric binary matrix factorization model to detect overlapping communities. The proposed method can assign community memberships to graph nodes, particularly; the outlier can also be distinguished from the overlapping nodes. Furthermore, this paper defines an improved partition density to evaluate the performance of community detection algorithm[22].

Fu et al. present a novel approach that utilizes belief propagation and conflict to detect communities. Firstly, they find triangles with maximal clustering coefficients as seed nodes. Secondly, the beliefs propagate their importance along the graph to construct their territory, and then conflict with each other when meeting the same node at the same time. Thirdly, the node membership is calculated by the belief vectors[23].

However, the above researches have not considered the user influence in community mining, and the weighted network model has not been used yet. Furthermore, the relationships between users and the membership degree of users belonged to a community have not fully taken into account in former related works. Different from the above works, in this paper, we proposed a novel overlapping community mining algorithm fully utilizing the above factors which have not been considered in existing works.

3 Problem description

The problem of micro-blog platform community mining can be modelled to graphs, where nodes represent users and edge denotes the relationship between users. Particularly, the weighted graph can be used to describe the relationship degree between users, and the user importance can also be considered in community mining problem.

Assuming that there is a network $G=(V,E)$, where V and E represent the vertex set and edge set respectively. Supposing that the community set is represented as $C=\{c_1,c_2,\cdots,c_k\}$, where $c_1\cup c_2\cup\cdots\cup c_k=V$, and $\exists i,j$ such that $c_i\cap c_j\neq\varnothing$. The aim of overlapping community mining is to seek the community set C. Some important definitions related to the problem of overlapping community mining are given as follows.

Definition 1: (Disjoint community structure) A disjoint community structure of graph $G=(V,E)$ is a partition of V into a set $P=\{C_1,C_2,\cdots,C_m\}$ of m non-empty subsets of vertex set V, such that, each node $u\in V$ is belonged to one of the following subsets:

1) The union of the node in set P is equal to V, that is, $\bigcup P=\bigcup_{i=1}^{m}C_i=V$.

2) The intersection of each two node in P is empty, that is, $\forall 1\le i<j\le m, C_i\cap C_j=\varnothing$.

Definition 2: (Overlapping community structure) An overlapping community structure of the graph $G=(V,E)$ refers to a cover of vertex V into a set Q of n non-empty subsets of V, such that the nodes of Q are covering $V:\bigcup Q=\bigcup_{i=1}^{n}C_i=V$.

Definition 3: (Membership degree) for a given community c and a user V_i, the membership degree between community c and user V_i is defined as

$$M(c,V_i)=\frac{1}{UI(V_i)}\cdot\sum_{V_j\in c}w_{ij} \tag{1}$$

Where $UI(V_i)$ means the user influence of V_i.

In Fig.1, we provide an example of overlapping community.

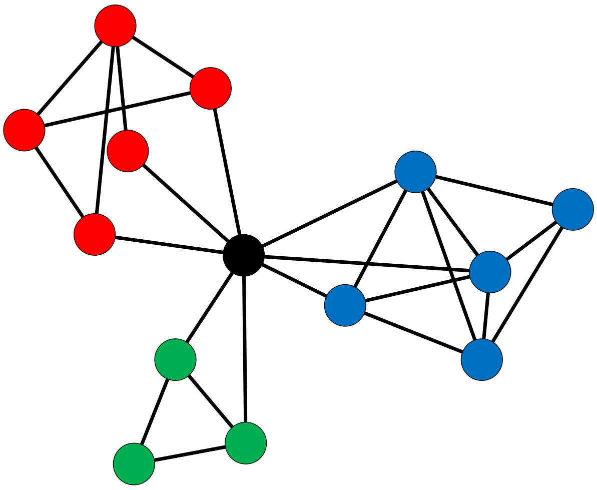

FIGURE .1 an example of overlapping community

As is shown in Fig. 1, the proposed network is made up of three communities, all of them are cliques. Particularly, all these communities are overlapping in the central node (represented as a black node), and these three communities (denoted as red, green, and blue nodes respectively) are densely intra-connected. Moreover, they are not sparsely connected with other communities. Hence, new community definition should be illustrated for overlapping community mining for the micro-blog platform.

4 User influence calculating

In social networks, the influence of users and relationship between users are different. Hence, to model social networks, the weighted networks are more suitable. Definitions of weighted networks are given as follows.

For the given vertices V_i and V_j, the edge weight E_{ij} is define as follows.

$$EW_{ij}=\begin{cases} w_{ij},\ if\ V_i\ and\ V_j\ are\ connnected\ in\ the\ weighted\ network \\ 1,\ if\ V_i\ and\ V_j\ are\ connnected\ in\ the\ unweighted\ network \\ 0,\ otherwise \end{cases} \tag{2}$$

The vertex weight VW_i of vertex V_i is defined as:

$$VW_i=\begin{cases} \sum_{V_j\in V} E_{ij},\ if\ V_j\ is\ the\ neighbor\ nodes\ of\ V_i \\ 0,\ otherwise \end{cases} \tag{3}$$

As is well known, each user in the micro-blog platform can promote the prestige by publishing micro-blog, forwarding and commenting blog of other users. The attribute in the micro-blog mainly contains two aspects: 1) user attributes and 2) micro-blog properties. User attributes contain user ID, user type, attention number, number of fans, number of micro-blog, number of mentions and so on. On the other hand, micro-blog attributes include number of micro-blog; publish date and time, the forwarding numbers, numbers of comments and so on. Particularly, node's attribute in social network belongs to basic characteristics of a node. If the user influence degree is correctly estimated, the quality of overlapping communities mining can be promoted obviously. Particularly, when estimating of user influence, the interaction behaviours between users should be considered carefully. Afterwards, the analytic hierarchy process is used to solve the weight of each influencing factor, and then the user influence can be calculated as the following steps:

(1) Constructing the index matrix X, and then normalized it.

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{4}$$

Afterwards, utilizing the standard 0/1 transform, the following equation can be obtained.

$$a_{ij}=\frac{x_{ij}-x_j^{\min}}{x_j^{\max}-x_{ij}} \tag{5}$$

Where $x_j^{\max}$ and $x_j^{\min}$ refer to the max and min value in the j^{th} column respectively, and then a new can be obtained.

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{6}$$

(2) Utilizing the analysis hierarchy process technology to compute the weight of index.

(3) Computing the weight of user attribute as follows.

$$\psi(V_i)=\sum_{j=1}^{m} ai_j\cdot w_j,\ i\in\{1,2,\cdots,n\} \tag{7}$$

(4) Computing the weight of user interactive behavior as follows.

$$\gamma(V_i)=(1\text{-}\lambda)\cdot\gamma(V_i)+\lambda\cdot\sum_{V_j\in N(V_i)} p_{ji}\cdot\gamma(V_j) \tag{8}$$

Where p_{ji} means the distribution probability, $N(V_i)$ refers to the set of nodes which points to the node V_i in the weighted graph, and λ denote the parameter which is ranged in $[0,1]$.

$$p_{ji}=\frac{N_{ji}}{\sum_{y\in O(j)} p_{jy}} \tag{9}$$

Where N_{ji} means the number of node j forward node i, $O(j)$ represents the set of node j forwards. N_{ji} Is the number of j forwards others, and p_{ji} means the probability of node j forward node i.

Afterwards, for the given user V_i, its influence can be obtained by the following linear fusion process.

$$I(V_i)=\eta\cdot\psi(V_i)+(1\text{-}\eta)\cdot\gamma(V_i),\ i\in\{1,2,\cdots,n\} \tag{10}$$

Where η is the regulatory factor. If the user attribute is more important than the interactive behaviour, η is set to larger than 0.5, otherwise, η is set to smaller than 0.5.

5 Overlapping community mining algorithm based on weighted network

The proposed overlapping community mining algorithm is designed based on weighted network; hence, weighted network construction is an important step in this paper. To build a weighted network, we should construct an unweighted network with a specific topological mixing parameter μ_t, and then allocate a positive real number to each edge of the graph.

Afterwards, parameter β should be allocated a weight s_i to each node. Another parameter μ_w is utilized to allocate a weight $s_i^{in}=(1-\mu_w)\cdot s_i$, which refers to the sum of weights of the edges between node V_i and all its neighbor nodes. Our overlapping community mining algorithm is given as follows.

Algorithm 1:
Overlapping community mining algorithm based on weighted network

```
Input: Weighted graph $G = (V, E)$,
Output: Community mining results
(1) Constructing the original nodes set $O = V$
(2) For each node ($V_i$), computing its node weight (that
    is, user influence) as follows.
    $W(V_i) = \sum_{j \in O} w_{ij}$                                    (11)
    Where $w_{ij}$ refers to the weight of edge $E_{ij}$
(3) Choosing the user with highest influence, and find
    this user's neighbors from the set $O$.
(4) Using the users chosen from step 3 to construct the
    initial community $c_0$
(5) For each user in the community $c_0$
(6) If the membership degree of user $i$ to the
    community $c_0$ is smaller than a threshold (in this
    paper, we set it to 0.5)
(7) Removing user $i$ from community $c_0$
(8) End if
(9) End For
(10) Repeating step 5 to step 9, until $\forall\ j \in c_0$, and the
    membership degree of user $j$ to the community $c_0$ is
    equal or larger than the pre-defined threshold.
(11) Searching all neighbors of the community $c_0$
    (denoted as $NB(c_0)$)
(12) Computing the membership degree $M(k, c_0)$ for
    each neighbor $k$
(13) Searching the users who can satisfied the following
    two conditions:
    Condition 1: $M(k, c_0) > \chi_2$
    Condition 2: $\chi_2 \geq M(k, c_0) \geq \chi_1$
    Where $\chi_1$ and $\chi_2$ represent two pre-defined
    thresholds. (In this experiment, $\chi_1$ and $\chi_2$ are set to
    0.4 and 0.5 respectively.
(14) Computing $N_1 = \{k | M(k, c_0) > \chi_2\}$
(15) Computing $N_2 = \{k | \chi_2 \geq M(k, c_0) \geq \chi_1\}$
(16) If $|N_1| > 0$
(17) Put all the users of $N_1$ into the community $c_0$, go to
    step 11.
(18) End if
(19) If $|N_2| > 0$
(20) For each user $V_i$ in $N_2$, put it into the community
    $c_0$ if the modularity can be higher after this
    operation, go to step 11.
(21) End if
(22) If $|N_1| = 0$ and $|N_2| = 0$ are satisfied
(23) Output the current community detection states as the
    final overlapping community mining results.
```

6 Experiment

6.1 DATASET

In this experiment, to make the performance evaluation more objective, some standard datasets are chosen (shown in Table.1). Particularly, to testify the effectiveness of the proposed algorithm on micro-blog platform, we construct a micro-blog dataset that is collected by Twitter API.

TABLE1 Standard datasets used for overlapping communities mining.

Dataset name	Number of Nodes	Number of Edges
Football[24]	115	613
Dolphins[25]	62	159
PGP[26]	10680	24340
Soc-epinions[27]	75879	508837
CA-GrQc[28]	5242	28980

Apart from the above standard dataset, a real dataset collected from micro-blog platform (twitter) is constructed. Twitter provides two application programming interfaces (APIs) to automatically crawl tweets, that is 1) search API, that can retrieve past tweets matching a user specified criteria and 2) streaming API that can subscribe to a continuing live stream of new tweets matching a user defined criteria.

We used the Twitter dataset collected from English-speaking users for evaluation. We used Twitter REST API to facilitate the data collection. The majority of the tweets collected were generated in a month period from July 4, 2013 through August 3, 2013. To prune incomplete and noisy twitter data, text pre-processing is utilized in advance. Five steps pre-processing are implemented, including: 1) discarding retweeted tweets, 2) deleting tweets with less than 6 words, 3) removing tweets with non-English words and 4) removing stop-words, URLs, user names, 5) stemming all the rest words. Finally, after implementing the above steps, 90.7M tweets are obtained, and this dataset is named "Twitter dataset".

6.2 PERFORMANCE EVALUATION METRIC

In this experiment, NMI, F1 score, and Modularity are utilized as the performance evaluation metric. NMI refers to an information theoretic based metric, and it can evaluate the quality of clusters. The formal definition of NMI is defined as follows.

$$NMI = -\frac{2 \cdot \sum_{ij} N_{ij} \cdot \log\left(\frac{N_{ij} N}{N_i N_j}\right)}{\sum_i N_i \cdot \log\left(\frac{N_i}{N}\right) + \sum_j N_j \cdot \log\left(\frac{N_j}{N}\right)}. \qquad (12)$$

Where N represents the confusion matrix, and N_{ij} denotes the number of nodes in cluster X_i and Y_j. Next, N_i and N_j refer to the sum over row i and column j of confusion matrix N. F1 score is the metric combining precision and recall, and its definition is illustrated as follows.

$$F_1 = 2\frac{P \cdot R}{P + R}. \tag{13}$$

Where symbols P and R denote precision and recall respectively.

For evaluating the quality of network partitions, Newman and Girvan proposed the modularity measure Q [29] which has been widely used in quality evaluation of community discovery.

$$Q = \sum_{i=1}^{c}\left[\frac{A(V_i,V_i)}{A(V,V)} - \left(\frac{A(V_i,V)}{A(V,V)}\right)^2\right] \tag{14}$$

Where $A(V_i,V_j) = \sum_{u \in V_i, v \in V_j} k_{u,v}$ and $k_{u,v}$ means the weight of edge E_{uv}.

6.3 EXPERIMENTAL RESULTS AND ANALYSIS

To objectively make performance evaluation, several existing related works are chosen to compare with our proposed algorithms, which are SLPA [30], OSLOM [31], Game [32], COPRA [33]. Xie et al. proposed a novel overlapping munity detection algorithm based on label propagation (named SLPA). Main ideas of this method lie in that adjacent nodes exchange labels based on some pre-determined rules [30]. OSLOM (Order Statistics Local Optimization Method) is designed utilizing the optimization of a fitness function. As is well known that OSLOM can be regarded as a standalone independent algorithm for overlapping community detection [31]. Game represents a novel algorithm for overlapping community detection using concepts from game theory, hence, named Game[32]. COPRA refers to an extension of the label propagation algorithm, that is suitable to be used in overlapping community detection. In COPRA, each node can update its coefficients through averaging the coefficients of neighbors[33]. Particularly, to let other algorithms can achieve their best performance, we set the optimal parameters for them before this experiment. In the SLPA algorithm the parameter r is ranged in $[0.05, 0.5]$ with the increasing interval 0.05. For the COPRA the range of parameter v is between 1 and 9. For our proposed algorithm, the mixing parameter is set to be 0.1.

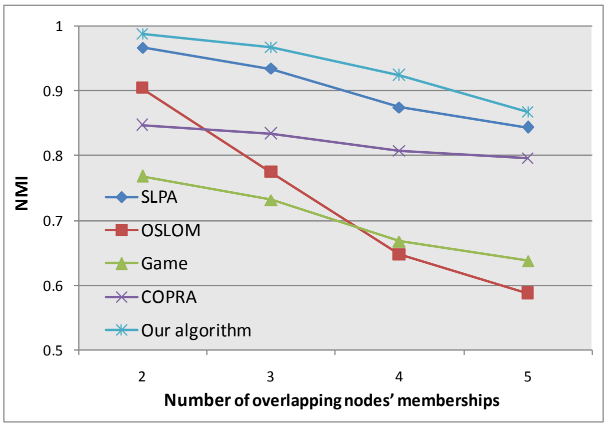

FIGURE 2 Performance evaluation using CA-GrQc under NMI metric

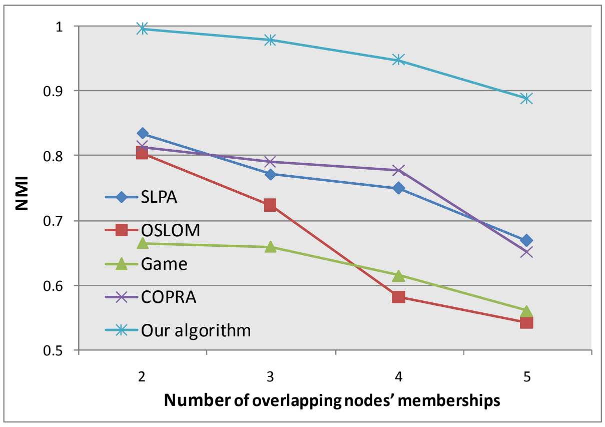

FIGURE 3 Performance evaluation using Twitter dataset under NMI metric

In Fig.2 and Fig.3, NMI metric is used to make performance evaluation, and dataset CA-GrQc and Twitter are utilized. Next, we make performance comparison for five methods (that is, SLPA, OSLOM, Game, COPRA, Our algorithm) with the number of overlapping nodes' memberships changing. Experimental results in Fig.2 and Fig.3 show that our algorithm performs better than other methods for NMI, and the proposed algorithm is more suitable to be used in the Twitter dataset. It means that 1) our algorithm can effectively solve different topological structures of social networks, 2) considering the factor of user influence, using our proposed algorithm, the performance of overlapping algorithm mining can be enhance greatly.

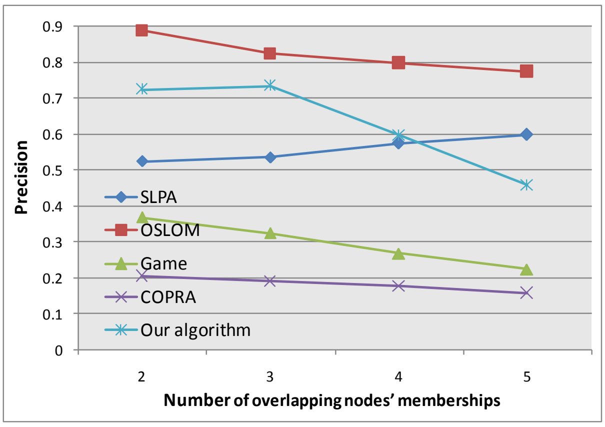

FIGURE 4 Performance evaluation using CA-GrQc under Precision metric

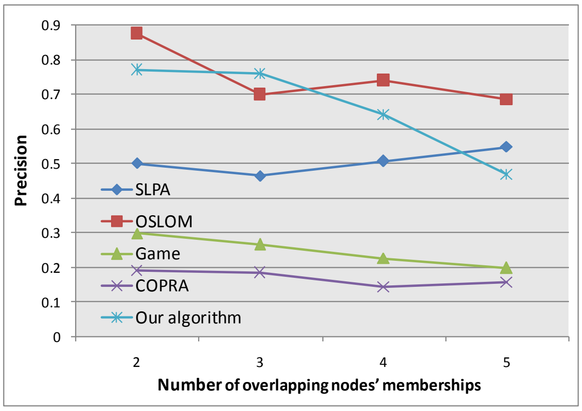

FIGURE5 Performance evaluation using Twitter dataset under Precision metric

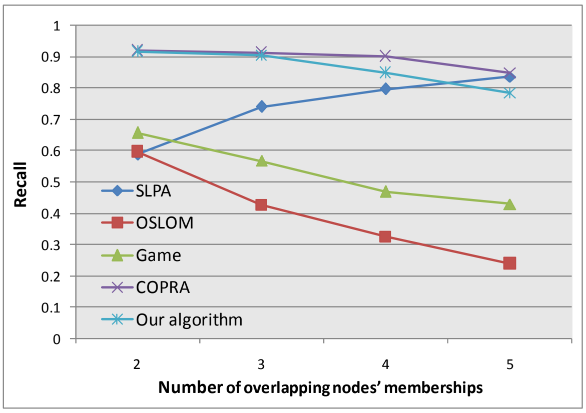

FIGURE 6 Performance evaluation using CA-GrQc under Recall metric

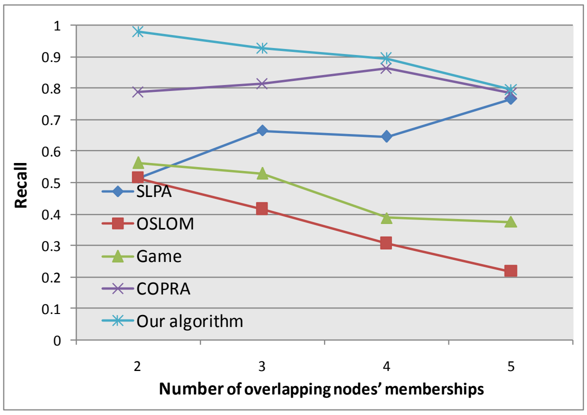

FIGURE 7 Performance evaluation using Twitter dataset under Recall metric

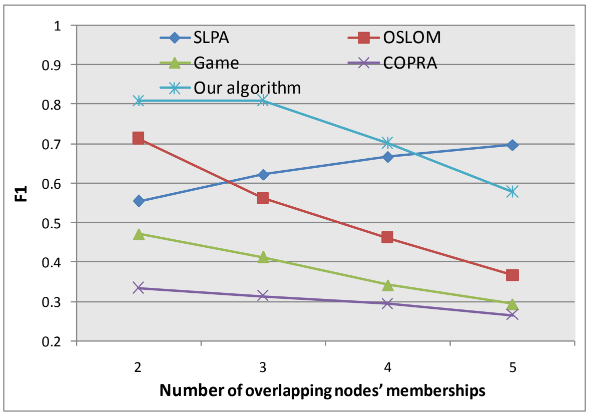

FIGURE 8 Performance evaluation using CA-GrQc under F1 metric

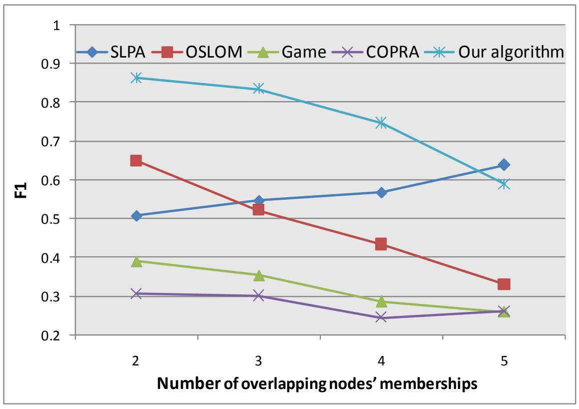

FIGURE 9 Performance evaluation using Twitter dataset under F1 metric

From Fig.4-Fig.9, to test the accuracy overlapping nodes recognition, Precision, Recall and F1 measure are utilized for these five methods with the number of overlapping nodes' membership varying. The task of overlapping nodes recognition is just like a binary classification problem. The Precision metric is calculated through dividing the number of overlapping nodes detected correctly by the total number of all detected overlapping nodes. On the other hand, the Recall metric is obtained through dividing the number of overlapping nodes detected correctly by the number of overlapping nodes which exist in the networks actually. Integrating the experimental results in Fig.4-Fig.9, it can be seen our algorithm outperforms other methods, especially on the Twitter dataset.

Combining all the experimental results above, the average values of all the above experiments are listed in Table.2, and the max value of each experiment is marked as bold.

TABLE.2 Performance comparison for different methods under different metrics

Metric	Dataset	SLPA	OSLOM	Game	COPRA	Our algorithm
NMI	CA-GrQc	0.905	0.728	0.701	0.821	**0.936**
	Twitter dataset	0.756	0.662	0.625	0.758	**0.952**
Precision	CA-GrQc	0.558	**0.821**	0.296	0.181	0.628
	Twitter dataset	0.505	**0.750**	0.246	0.168	0.660
Recall	CA-GrQc	0.740	0.397	0.530	**0.896**	0.863
	Twitter dataset	0.648	0.363	0.464	0.811	**0.899**
F1	CA-GrQc	0.635	0.525	0.380	0.301	**0.724**
	Twitter dataset	0.565	0.483	0.322	0.277	**0.758**

As is shown in Table. 2, our algorithm performs best in NMI and F1, and in other metrics (such as Precision and Recall) our algorithm is quite effective as well.

Afterwards, to go a step further to evaluate the performance and precision of overlapping community mining in real-world social networks, the modularity metric is utilized and the experimental result is shown in Fig.10. The value of modularity highly relies on the number of overlapping communities to which each user belongs and the membership degree to each overlapping community. Particularly, we suppose that user belongs equally to all of the communities.

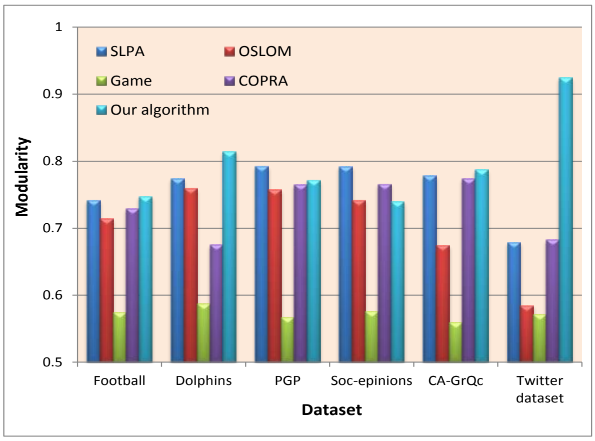

FIGURE 10 Performance evaluation using different dataset under Modularity metric

Fig.10 shows that for some networks like Soc-epinions, PGP, and Football, the results of SLPA are slightly higher than our algorithm, because the performance of our algorithm is not good when the value of the strictness is high. On the other hand, when the edges of the input network can represent relationships between different users, the proposed algorithm can perform better than other methods. The reason lies in that, in our algorithm, the modularity measure can provide importance to the topological properties of social networks. Particularly, the performance of our algorithm using Modularity is obviously better than others, because our algorithm utilize the user influence to construct weighted network, which can effectively model the overlapping community mining problem on micro-blog platform.

6.4 TIME COST OF THE PROPOSED ALGORITHM

The above experimental results demonstrate the effectiveness of the proposed algorithm, however, time cost of the proposed algorithm is another important problem. Hence, in this sub-section, time cost of our algorithm is analyzed. The detecting overlapping community algorithm proposed in this paper is implemented using Java programming language, and it is executed on a PC with 2.67 GHz processor, 4 GB memory and Windows 8 OS. As is shown in Fig. 10, a graph with 100 nodes is constructed with the density from 0 to 1 and the increasing interval is set to 0.001. We can see that all nodes are disconnected when density equals 0, and the graph can be regarded as a large clique if graph density is equal to 1.

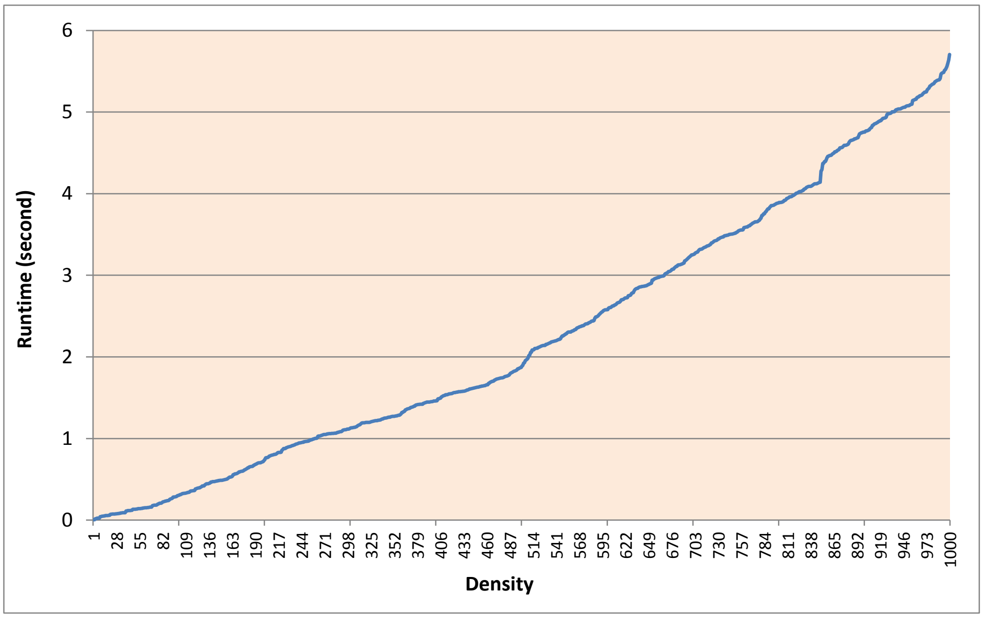

FIGURE 11 Runtime for the proposed algorithm with node density varying

From Fig.11, conclusions can be drawn that our proposed algorithm can effectively solve the graph with high density, because the speed of runtime increasing is almost linearly. Therefore, this proposed algorithm suitable to be performed in parallel.

In a word, the above experiments demonstrate the effective of our proposed algorithm in overlapping community mining. The reasons lie in the following aspects:

(1) We use the weighted network to model the problem of community mining in micro-blog platform.

(2) User influence in introduced in overlapping community mining, and this factor can effectively describe the internal structure of the micro-blog platform like social network.
(3) Our overlapping community mining algorithm can find the original communities in advance, and can effectively expand them to obtain the final results
(4) Our proposed algorithm is suitable to be implemented in parallel; hence, our algorithm can effectively solve the large-scale overlapping community mining problem.

7 Conclusions and future works

In this paper, we present a novel overlapping community mining algorithm for micro-blog platform. Particularly, we convert the problem of overlapping community mining to a weighted graph computation. Particularly, in the proposed algorithm user influence is represented as an important factor in overlapping community mining. In our algorithm, the original overlapping communities are detected in advance and then community mining results are gained by expanding the original communities. In the future, we will extend this work in the following aspects:
(1) We will expand the experiment dataset using other languages, such as Chinese, French, and Portuguese and so on.
(2) Some other factors will be considered in community mining, such as the dynamical user interest and the geographical position of users.
(3) Apart from the micro-blog platforms, we will testify if the proposed algorithm is suitable to be used in other types of social networks, such as Facebook, Youtube, Flickr and so on.

References

[1] Xie Jierui, Kelley Stephen, Szymanski Boleslaw K., Overlapping Community Detection in Networks: The State-of-the-Art and Comparative Study, *ACM Computing Surveys*, vol. 45, no. 4, Article No. 43, 2013.
[2] Allen, S. M.; Chorley, M. J.; Colombo, G. B., Exploiting user interest similarity and social links for micro-blog forwarding in mobile opportunistic networks, *Pervasive and Mobile Computing*, vol. 11, pp. 106-131, 2014.
[3] Yang Dong-Hui, Yu Guang, Static analysis and exponential random graph modelling for micro-blog network, *Journal of Information Science*, vol. 40, no. 1, pp. 3-14, 2014.
[4] Liu Ying, Moser Jason, Aviyente Selin, Network Community Structure Detection for Directional Neural Networks Inferred From Multichannel Multisubject EEG Data, *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 7, pp. 1919-1930, 2014.
[5] Zhang Zhiyuan, Feng Xia, Huo Weigang, An improvement of PLSA-based community detection algorithm, *Modern Physics Letters B*, vol. 28, no. 15, Article No. 1450120, 2014.
[6] Jia Songwei, Gao Lin, Gao Yong, Anti-triangle centrality-based community detection in complex networks, *IET Systems Biology*, vol. 8, no. 3, pp. 116-125, 2014.
[7] Qiu Jiangtao, Lin Zhangxi, D-HOCS: an algorithm for discovering the hierarchical overlapping community structure of a social network, *Journal of Intelligent Information Systems*, vol. 42, no. 3, pp. 353-370, 2014.
[8] Wang Zhu, Zhang Daqing, Zhou Xingshe, Discovering and Profiling Overlapping Communities in Location-Based Social Networks, *IEEE Transactions on Systems Man Cybernetics-systems*, vol. 44, no. 4, pp. 499-509, 2014.
[9] Li H. J., Zhang, J., Liu Z. P., Identifying overlapping communities in social networks using multi-scale local information expansion, *European Physical Journal B*, vol. 85, no. 6, Article No. 190, 2012.
[10] Adebayo Julius, Musso Tiziana, Virdee Kawandeep, An Exploration of Social Identity: The Structure of the BBC News-Sharing Community on Twitter, *COMPLEXITY*, vol. 19, no. 5, pp. 55-63, 2014.
[11] Pearce Warren, Holmberg Kim, Hellsten Iina, Climate Change on Twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group 1 Report, *PLOS ONE*, vol. 9, no. 4, 2014.
[12] Herdagdelen Amac, Zuo Wenyun, Gard-Murray Alexander, An Exploration of Social Identity: The Geography and Politics of News-Sharing Communities in Twitter, *COMPLEXITY*, vol. 19, no. 2, pp. 10-20, 2013.
[13] Ikeda Kazushi, Hattori Gen, Ono Chihiro, Twitter user profiling based on text and community mining for market analysis, *Knowledge-based Systems*, vol. 51, pp. 35-47, 2013.
[14] Kim Yong-Hyuk, Seo Sehoon, Ha Yong-Ho, Two Applications of Clustering Techniques to Twitter: Community Detection and Issue Extraction, *Discrete Dynamics in Nature and Society*, Article No. 903765, 2013.
[15] Rhouma Delel, Ben RomdhaneLotfi, An efficient algorithm for community mining with overlap in social networks, *Expert Systems with Applications*, 2014, 41(9): 4309-4321.
[16] Chen Y., Wang X. L., Yuan B., Overlapping community detection in networks with positive and negative links, *Journal Of Statistical Mechanics-theory and Experiment*, Article No. P03021, 2014.
[17] Cui Laizhong, Qin Lei, Lu Nan, A Fast Overlapping Community Detection Algorithm with Self-Correcting Ability, *Scientific World Journal*, Article No. 738206, 2014.
[18] Wang Wenjun, Liu Dong, Liu Xiao, Fuzzy overlapping community detection based on local random walk and multidimensional scaling, *Physica A-statistical Mechanics and Its Applications*, vol. 392, no. 24, pp. 6578-6586, 2013.
[19] Wang Zhenwen, Hu Yanli, Xiao, Weidong Overlapping community detection using a generative model for networks, *Physica A-statistical Mechanics and Its Applications*, vol. 392, no. 20, pp. 5218-5230, 2013.
[20] Badie Reza, Aleahmad Abolfazl, Asadpour Masoud, An efficient agent-based algorithm for overlapping community detection using nodes' closeness, *Physica A-statistical Mechanics and Its Applications*, vol. 392, no. 20, pp. 5231-5247, 2013.
[21] Shi Chuan, Cai Yanan, Fu Di, A link clustering based overlapping community detection algorithm, *Data & Knowledge Engineering*, vol. 87, pp. 394-404, 2013.
[22] Zhang Zhong-Yuan, Wang Yong, Ahn Yong-Yeol, Overlapping community detection in complex networks using symmetric binary matrix factorization, *Physical Review E*, vol. 87, no. 6, Article No. 62803, 2013.
[23] Fu Xianghua, Liu Liandong, Wang Chao, Detection of community overlap according to belief propagation and conflict, Physica A-statistical Mechanics and Its Applications, vol. 392, no. 4, pp. 941-952, 2013.
[24] M. Girvan, M.E.J. Newman, Community structure in social and biological networks, *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821-7826, 2002.
[25] D. Lusseau, K. Schneider, O.J. Boisseau, P. Haase, E. Slooten, S.M. Dawson, The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations, *Behavioral Ecology and Sociobiology*, vol. 54, pp. 396-405, 2003.
[26] Boguna M, Pastor-Satorras R, Diaz-Guilera A, Models of social networks based on social distance attachment, *Physical Review E*, vol. 70, no.5, Articule No. 056122, 2004.
[27] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, *in: The Semantic Web-ISWC 2003*, pp. 351-368, 2003.
[28] J. Leskovec, J. Kleinberg, C. Faloutsos, Graph evolution: densification and shrinking diameters, *ACM Transactions on Knowledge Discovery from Data*, vol.2, pp.1-42, 2007.

[29] Newman M E J, Girvan M. Finding and evaluating community structure in networks. *Physical Review E*, vol. 69, no.2, Article No. 026113, 2004.

[30] Xie Jierui, Szymanski Boleslaw K, Towards linear time overlapping community detection in social networks, *Advances in Knowledge Discovery and Data Mining*, pp.25-36, 2012.

[31] A. Lancichinetti, F. Radicchi, J.J. Ramasco, S. Fortunato, Finding statistically significant communities in networks, *PLoS One*, vol. 6, Article No. e18961, 2011.

[32] W. Chen, Z. Liu, X. Sun, Y. Wang, A game-theoretic framework to identify overlapping communities in social networks, *Data Mining and Knowledge Discovery*, vol. 21, no.2, pp.224-240, 2010.

[33] U.N. Raghavan, R. Albert, S. Kumara, Near linear time algorithm to detect community structures in large-scale networks, *Physical Review E*, vol. 76, Article No. 036106, 2007.

Author

<Zhaoyin Zhang>, <1955.10>,< Harbin County, Heilongjiang Province, P.R. China >

Current position, grades: the Associate professor of School of, Heilongjiang University, China.
University studies: received his B.Sc. in Electrical Engineering and Automation from Heilongjiang University of Heilongjiang in China. He received his M.Sc. from Heilongjiang University in China.
Scientific interest: His research interest fields include: Computer application; Artificial intelligence; Robot; Software engineering.
Publications: more than fifteen papers published in various journals.
Experience: He has teaching experience of thirty-two years, has completed three scientific research projects.