Send Orders for Reprints to reprints@benthamscience.ae

1188 *The Open Cybernetics & Systemics Journal,* 2014, *8,* 1188-1197

Open Access

Community Detection of Chinese Micro-Blogging Using Multi-Dimensional Weighted Network

Xiaoping Zhou[1,2,*], Xun Liang[1] and and Run Cao[1]

[1]*Renmin University of China, Beijing 100782, China;* [2]*Beijing University of Civil Engineering and Architecture, Beijing, 100040, China*

Abstract: Existing community detection methods are mostly based on the analysis of the links among the nodes, ignoring the rich, while the others often ignore the network structure which is the foundation of social media. Aiming at the existed problems, this paper proposed a community detection algorithm based on multi-dimensional weighted network. By introducing User Interactive Frequency, User Interest Similarity, and User Attributes Similarity into the basic network topology, a multi-dimensional weighted network is set up. After converting the multi-dimensional weighted network into a single-dimensional weighted network, an improved CNM algorithm is exploited to discover the communities. A corresponding series of evaluation indicators are proposed to evaluate the detection results. By evaluating the algorithm in the dataset of Chinese Micro-blogging, it reveals that the clustering results are better when extra information is used, and in Chinese Micro-blogging platform, User Interactive Frequency plays a much more important role in community detection.

Keywords: Chinese, CNM, community detection, micro-blogging, multi-dimensional, weighted network.

1. INTRODUCTION

Community can refer to a usually small, social unit of any size that shares common values. Since the advent of the Internet, the concept of community has less geographical limitation. Micro-blogging is a broadcast medium in the form of blogging. With hundreds of millions of people gathering virtually in a Micro-blogging, such as Facebook and sina Weibo (the most widely used Chinese Micro-blogging), it has been a typical online virtual community.

With the development of 3G technology and the gradual improvement of mobile access to the Internet, both the user scale and social influence of Micro-blogging sites have experienced rapid growth. Currently, Micro-blogging is becoming a widely-accepted platform to express people's views and opinions. By its application in communication, marketing, political purposes and so on, Micro-blogging plays important roles in our everyday life. Thus, the research results of community detection on Micro-blogging platform have practical significance, which can be used to guide the network marketing, public opinion control and many other issues. Essentially, for its hybrid network and the aspects of both social networks and information networks, Micro-blogging network is a complex network.

Community detection, which is fundamental for further researches on community in complex network, has gained of paramount attentions. As a representative complex network, Micro-blogging attracts a number of researches, some of which are working on community detection.

Existing community detection methods are mostly based on the analysis of the links between nodes and ignore the rich information [1-4] of new era social media, while the others [5, 6] often ignore the network structure which is the foundation of social media. With the continuous expansion of the research on complex network, especially on online communities, some scholars tried to apply the content of nodes and edges of the network to the community detection. Yan *et al*. proposed a community detection method, which integrated the interest and network topology [7]. Java *et al*. synthesized the links and the labels in folksonomy, proposed a novel community detection method, and got a more precise community classification [8]. Zhang *et al.* combined user friendship network and user-generated content, used NMF and author-topic AT model, and gained a better result in Delicious and Twitter dataset, thus proposing a method to detect the interest-based community by finding out the leading user in certain interest field.

The Micro-blogging contains a variety of links, vast amounts of user generated content, and a large number of user attribute metadata information which can be used for user clustering. Though, these data are from different dimensions, they are substantial for Micro-blogging platform which has sociability and media. And in Micro-blogging community detection, all of the related information should be taken into consideration.

This paper aims to introduce the related information into the basic Micro-blogging network topology, and gets a better community detection result.

The rest of the paper is organized as follows. The literature review is presented in Section 2. Section 3 presents in detail the problem definition and the framework. Section 4 introduces the detailed algorithm. Section 5 introduces a

1874-110X/14 2014 Bentham Open

Table 1. Comparison of classic community detection algorithms.

Algorithms	Time Complexity	Pros	Cons
Spectral Bisection Method	$O(n^3)$	Fast	Only divide the network into two parts
Kemighan-Lin Algorithm	$O(tn^2)$	Fast	Priori knowledge (*e.g.*, the number or the average scale of the communities) is required, and can't be apply to real network
W-H Algorithm	$O(n)$	Can detect which community the specific node belongs to, without the details of the whole community network structure	Partial information of the network structure is required
GN Algorithm	$O(m^2n)$	The number and the scale of division communities is not required, and precise	The evaluation of the reasonable number of division communities is not given, and high time complexity
Newman Algorithm	$O((m+n)n)$	Fast	Suboptimum solution
CNM Algorithm	$O(nlog^2n)$	Close to linear in time complexity, and space complexity is shortened by using the heap structure	Suboptimum solution

corresponding series of evaluation indicators. Experiments are presented and analyzed in Section 6. Finally, Section 7 concludes the paper with a summary of the future work.

2. LITERATURE REVIEW

In recent years, community detection has experienced a rapid development, and becomes significant research field in data mining. Though, a mass of researches and extensive discuss had launched; till now, there's no common method to definite a reasonable community structure of a given network.

Currently, community detection methods on Micro-blogging can be divided into 2 groups: clustering algorithm and link-based algorithm.

2.1. Clustering Algorithm

Clustering algorithm, which requires information or properties of the network except for the space topology, consists of traditional clustering algorithm (such as K-Mean Algorithm, DBSCAN Algorithm, and so on), text clustering algorithm (such as Latent Dirichlet Allocation Model, Vector Space Model, TF-IDF Model), and the algorithms derived from the above.

2.2. Link-based Algorithm

In some complex networks, it's difficult or cost-too-much to acquire information or properties. In these scenarios, only can the basic links between the nodes be used, and the community detection methods include traditional graphic-based algorithm (such as Kernighan-Lin Algorithm), spectral analysis clustering algorithm (such as Spectral Bisection Method, Unnormalized Spectral Clustering, and Normalized Spectral Clustering), splitting algorithm (such as GN algorithm) and aggregation algorithm (such as CNM algorithm).

As the only handful resource is the relations among the nodes in the primary network, how to treat the relations respectively in community detection is becoming a hot research problem. Weighting the edge of the network, which is proposed by M. Girvan and M.E.J. Newman in GN algorithm, has been extensive used. In GN algorithm, the edge betweenness is introduced to evaluate the influence of edge in the network. By removing the edge with the largest edge betweenness iteratively, it divides a huge network into small ones. GN algorithm, which has been proved in community detection, is limited in the network with small scale for its high costs in calculating. To solve the issue, some other fast algorithms were proposed, such as Newman algorithm, CNM algorithm. Table **1** shows the comparison of classic community detection algorithms.

3. PROBLEM DEFINITION

The related information (*e.g.*, a variety of links, vast amounts of user generated content, and a large number metadata of user attribute information) in a Micro-blogging is taken into account to obtain a better community classification. And the terminologies related to the work should be defined, and the overall solution framework should be given.

3.1. Terminology Definition

Currently, the research on online community structure is mainly on the detection tragedy and method. Micro-blogging is an online community with huge number of participants. And it can be defined in relation network's way.

Definition 1. Micro-blogging Topology Graph

Micro-blogging Topology Graph can be defined as:

$G = (V, E, W)$

where G is a weighted digraph, V is the set of nodes, E is the set of directed edges, W is the set of the weights of the edges.

Define V as $V = \{v_1, v_2, ..., v_n\}$, and $v_i \in V$ represents a Micro-blogging user. Define E as $E=\{e_{ij}|e_{ij}=(v_i,v_j)$, and $e_{ij} \in E$ represents the relationship between user v_i and user v_j, and v_i, $v_j \in V$. Define W as $W=\{w(e_{ij})| \forall e_{ij} \in E\}$, and w_{ij} represents the stength of the relation e_{ij}.

In Chinese Micro-blogging (sina Weibo), one user can follow any other user. And the follower can receive all the content the followee generated. Based on the follow relationship, a basic network, called follower-followee network, is built. Moreover, the follower-followee network can subdivide to two sub-networks. Only when both user v_i follow user v_j and user v_j follows user v_i could the relation e_{ij} is established. In this situation, the network is defined as Micro-blogging Bi-followed Network. In some other scenarios where the relation e_{ij} is set either when user v_i follow user v_j or when user v_j follows user v_i, the network is called Micro-blogging Single-followed Network.

Both Micro-blogging Bi-followed Network and Micro-blogging Single-followed Network are basic Micro-blogging networks. As Micro-blogging Bi-followed Network has a stricter construction condition, it's a sub network of Micro-blogging Single-followed Network, and has a sparse relationship. Needless to say, the strength of relation can be added to either network.

Definition 2. User Interactive Frequency

In Chinese Micro-blogging (sina Weibo), a user can forward another user's micro-blog or mention some other users in his micro-blog. User Interactive Frequency is defined to show how many times one user forward or mention another one. And User Interactive Frequency can be defined as:

$F = \{f(e_{ij})| \forall e_{ij} \in E\}$

where $f(e_{ij})$ indicates the interaction between user v_i and user v_j. In both Micro-blogging Bi-followed Network or Micro-blogging Single-followed Network, $f(e_{ij})$ equals to the sum of the times user v_i mentioned or forwarded user v_j and user v_j mentioned or forwarded user v_i. Obviously, F is a type of the strength for relation e_{ij}, and it is a certain type of W.

Definition 3. User Interest Similarity

To a certain extent, what a user published in his micro-blog reflects what he is interesting in. And by analyzing the content a user generated, the User Interest Model can be calculated. Define the users' interest model as:

$Interest_i = \{Word_1, Word_2, ..., Word_n\}$

where $Interest_i$ means the User Interest Model of user v_i. Apparently, one user is always interesting in a series of things. And each keyword $Word_i$ implies one item he cares. Moreover, user places distinct emphasis on each keyword. In other words, each keyword has its own interest weight. And Word can be described as:

$Word_i = (Keyword_i, Weight_i)$

Thus, the problem is converted to how to find out the keywords and their weights.

Different from English content, where all the words are separated naturally by character 'space', Chinese content should be segmented to find out the words. In this paper, Chinese segmentation system ICTCLAS, which is developed by Chinese Academic of Science, is used to gain the keywords.

After the Chinese word segmentation, TF-IDF is exploited to calculate the weight of each keyword. By ranking the keywords in the order of their TF-IDF value, top 20% keywords are selected into the users' User Interest Model.

By analyzing the content published by user, User Interest Model is obtained. User Interest Similarity is defined to described how much is the similarity in interest between two users in a Micro-blogging. And it can be defined as:

$C = \{c(e_{ij})| \forall e_{ij} \in E\}$

where $c(e_{ij})$ means the interest similarity between user v_i and user v_j.

From each user's User Interest Model $Interest_i$, it can be treated as a vector in geometry. And Vector Space Model (VSM) is easily coming out to solve the problem. By using VSM, $c(e_{ij})$ can be simply calculated by the following formula.

$$c(e_{ij}) = sim(Interest_i, Interest_j)$$

$$= \frac{\sum_{k=1}^{n} Weight_{ik} Weight_{jk}}{\sqrt{\sum_{k=1}^{n} Weight_{ik}^2 \times \sum_{k=1}^{n} Weight_{jk}^2}}$$

where $Weight_{ik}$ means the weight of $Word_k$ in user v_i's User Interest Model. Only the same keywords in User Interest Model are calculated in the formula. Undoubtly, C is another typical W.

Definition 4. User Attribute Similarity

In Chinese Micro-blogging (sina Weibo), User's basic information includes his location (province and city), gender, number of fans, number of friends, registration time and so on. And all of these data are significant for community detection of Micro-blogging. The location (province and city) implies the user's geographic position; the number of fans shows the influence in the network; the number of blogs shows the user's activity; the registration time shows his level in Micro-blogging. And User Attribute Model can be described as:

$$Attr = \{gender, createdtime, followers, friends, biFollowers, statuses, verified, province, city\}$$

$Attr_i$ represents the User Attribute Model of user v_i. Similarly, User Attribute Similarity can be defined as:

$A = \{a(e_{ij})| \forall e_{ij} \in E\}$

where $a(e_{ij})$ means the interest similarity between user v_i and user v_j.

Then VSM is utilized to calculate the User Attribute Similarity.

where w_{ik} means the value of the corresponding attribute of $Attr_i$. Also, A belongs to W.

$$a(e_{ij}) = sim(Interest_i, Interest_j)$$

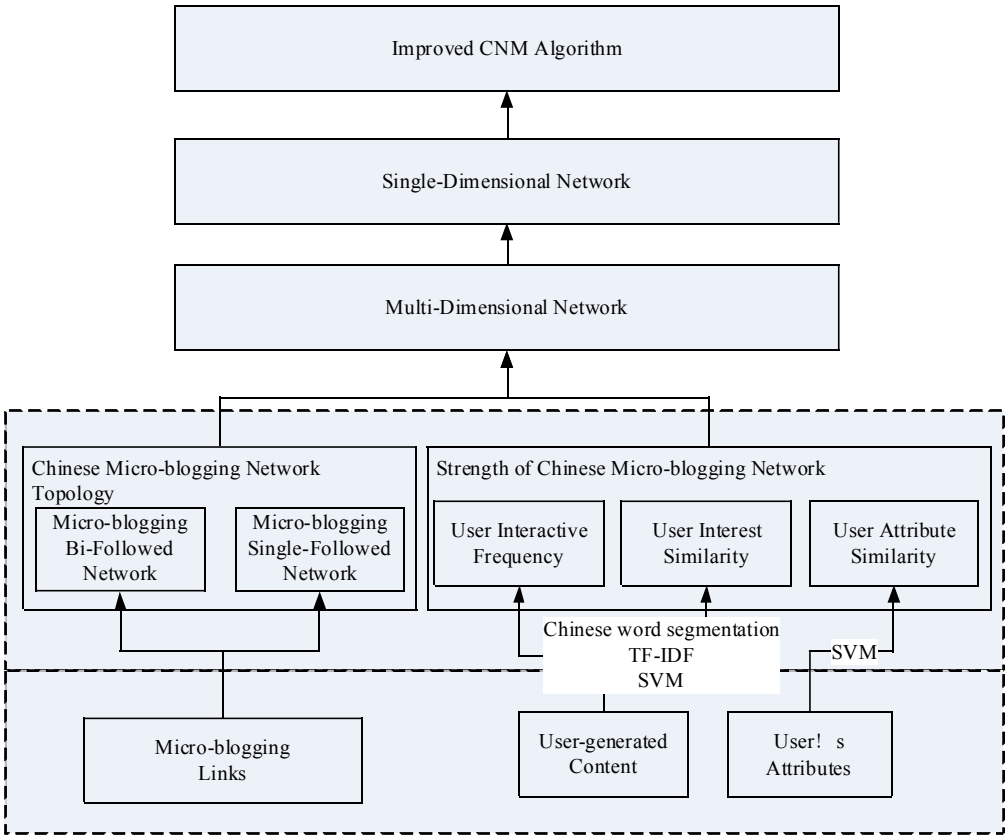

Fig. (1). Overall solution framework.

$$= \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \times \sum_{k=1}^{n} w_{jk}^2}}$$

3.2. Overall Solution Framework

As shown in Fig. (**1**), this paper builds the basic network topology through Micro-blogging links. It collects 7776 users with their attributes and user-generated content of sina Weibo. By analyzing the data, Micro-blogging Single-followed Network is formed, which is the basic for Micro-blogging Bi-followed Network. Users' attributes are used to calculate the User Attribute Similarity using SVM. User Interactive Frequency is extracted from the user-generated content. User Interest Similarity is worked out after three steps: Chinese word segmentation, TF-IDF and VSM.

Based on the basic network and the strength, multi-dimensional network is set up. And the community detection algorithm stated in section 4 is used to convert the multi-dimensional network into a single-dimensional network, where classic community detection algorithms can be used.

4. COMMUNITY DETECTION ALGORITHM

Aiming at the "complex" characteristics of Chinese Micro-blogging (sina Weibo), this paper proposes an community detection algorithm based on weighted networks. Combing the three typical edge betweennesses (User Interactive Frequency F, User Interest Similarity C, and User Attribute Similarity A) to either of the two basic Micro-blogging networks brings forth three single-dimensional networks. Each time when a new edge betweenness is added to the basic network, it generates a new single-dimensional network. As all these single-dimensional networks are based on the same basic network topology graphic, they can be treated as a multi-dimensional weighted network. On this point, the varieties of the information in Micro-blogging generate the multi-dimensional weighted network.

Considering a p-dimensional network as:

$$\widehat{A} = \{A^{(1)}, A^{(2)}, \cdots, A^{(p)}\}$$

where $A^{(i)} \in \{0,1\}^{n \times n}$ means the interactive status in the i-th dimension.

As there will exist kinds of potential interactive methods in a Micro-blogging, the underlying shared community structure will be extremely complicated, which make it necessary to integrate the multi-dimensional network to find out the shared community structure.

A valid quomodo to process the multi-dimensional network is to convert it into a single-dimensional network. From the perspective of set theory, a variety of interactive forms among nodes strengthen their relation. And a multi-dimensional network can be transformed to:

$$\overline{A} = \sum_{i=1}^{p} \theta_i A^{(i)}$$

where θ_i signifies the importance of the *i*-th dimension network in the *p*-dimensional network. And often, there exists:

$$\sum_{i=1}^{p} \theta_i = 1$$

Sometimes, θ_i is set to be *1/p*.

Using $\overline{A}$, community detection in multi-dimensional network is converted to a simple one-dimensional network, where the classic community detection algorithms can be applied. And an improved CNM algorithm is used in this paper.

5. EVALUATION INDICATORS

In order to meet the needs of specific scenario, kinds of community detection algorithms are founded. Meanwhile, a series of evaluation tragedies are formed to evaluate the detection results. In this section, a new evaluation tragedy is proposed correspondingly.

For any community, users and topics are two main elements. And the evaluation indicators for the detection result should consider both the users and topics. As Modularity *Q*, which is proposed by Newman in CNM algorithm, has been proved, it's also selected in the evaluation system and named *Qs*. Moreover, some other indicators related to users or/and topics are introduced.

Definition 5. User Interactive Frequency Indicator

User Interactive Frequency Indicator is defined to describe the level of activity during communities. A good community classification should result in active interaction during communities. And User Interactive Frequency Indicator is defined as:

$$Frequency = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{\gamma} R_{ki}R_{kj}f(e_{ij})}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{\gamma} R_{ki}R_{kj}}$$

where $R_{ki} = 1$ when user *i* belongs to the community *k*, otherwise $R_{ki} = 0$.

Undoubtedly, good community detection result has nice *Frequency*.

Definition 6. User Interest Similarity Indicator

The users in the same community should have many similar interests. And User Interest Similarity Indicator is proposed to indicate the similarity during communities. The definition is:

$$ContentSu = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{\gamma} R_{ki}R_{kj}c(e_{ij})}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{\gamma} R_{ki}R_{kj}}$$

where $R_{ki} = 1$ when user *i* belongs to the community k, otherwise $R_{ki} = 0$.

Definition 7. User Attribute Similartity Indicator

It is supported that most users have the same attributes, such as living in the same city, having the same age range, and so on. And User Attribute Similarity Indicator is introduced as below:

$$userSu = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{\gamma} R_{ki}R_{kj}a(e_{ij})}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{\gamma} R_{ki}R_{kj}}$$

where $R_{ki} = 1$ when user *i* belongs to the community *k*, otherwise $R_{ki} = 0$.

Definition 8. Comprehensive Indicator

In order to balance the density and similarity in divided communities, a Comprehensive Indicator $\overline{\mu}$ is used and calculated as the average of *Frequency, ContentSu, UserSu* and *Qs*.

$$\overline{\mu} = \frac{Q_s + Frequency + ContentSu + UserSu}{4}$$

In the evaluation system, a higher Qs indicates closer relationship among the users in the community, larger value of Frequency means a larger average number of interaction (mention and forward), bigger number of ContentSu shows the more similar interest of the users and the content, and greater UserSu presents the higher similarity of users' attributes in community. For a good community division, users should have closed relations and share more contents and interest. And it can be evaluated by the indicators stated above.

6. EXPERIMENTS

Based on the two basic networks (Micro-blogging Bi-Followed Network and Micro-blogging Single-Followed Network) and three weighted methods (User Interactive Frequency, User Interest Similarity, and User Attributes Similarity), 8 experiments, which are divided into 2 groups according to the basic network, are designed to evaluate the community detection results. The former 4 experiments are based on Micro-blogging Single-Followed Network, and the rest on Micro-blogging Bi-Followed Network.

All the experiments use the same data: the same users, the same user relations. The experiment primary network has 1371 users, selected from 7776 users from Chinese Micro-blogging (sina Weibo) who has at least 50 fans.

As shows in Table **2**, no extra information is introduced in Exp. 1, which means all the weights of the edges in Exp.1 are set 1; Exp. 2 uses User Interactive Frequency as the strength of the edges; Exp. 3 combines User Interest Similarity and User Attributes Similarity to the edges using multi-dimensional network algorithm stated above. Exp. 4 takes all the three factors into the consideration.

Table 2. Experiments based on micro-blogging single-followed network.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Network	Micro-blogging Single-followed Network			
Weight	1	UIF	UIS UAS	UIF UIS UAS
Algorithm	CNM	CNM	CNM	CNM
Output	division communities			

Table 3. Community detection results.

Indicator	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Qs	0.4532	0.5293	0.4728	0.4949
Frequency	1.2341	1.2798	1.2333	1.2529
ContentSu	0.0387	0.0396	0.0451	0.0445
UserSu	0.4134	0.4169	0.4148	0.4163
$\overline{\mu}$	0.5348	0.5664	0.5415	0.5521
Time(s)	114.345	114.427	116.874	110.610

Table 4. Experiments based on micro-blogging bi-followed network.

	Exp. 5	Exp. 6	Exp. 7	Exp. 8
Network	Micro-blogging Bi-followed Network			
Weight	1	UIF	UIS UAS	UIF UIS UAS
Algorithm	CNM	CNM	CNM	CNM
Output	division communities			

The experiments exploit the evaluation indicators stated above to estimate the detection results. Table **3** shows the detection results in Micro-blogging Bi-Followed Network.

Exp. 1, which no more extra information is used to set up the strength of the edges, has the lowest number in all evaluation indicators. It reveals that extra valuable data help to obtain better community division. Both Exp. 3 and Exp. 4 get a good mark in all the indicators. However, an interesting result is that Exp. 2 has the highest score in almost all the indicators except for the *ContentSu*. The $\overline{\mu}$ values 0.5664 in Exp. 2, which is largest in the 4 experiments. And in Exp. 4, which takes all the three factors into accounts, $\overline{\mu}$ scores less than Exp. 2, but better than Exp.1 and Exp. 3, with the number of 0.5521. It demonstrates that in Micro-blogging Single-Followed Network, it is not that the more information results in better community detection. Moreover, the experiments show that in Micro-blogging Single-Followed Network, User Interactive Frequency is key relation strength, and it can dramatically improve the community detection result.

Table **4** shows the experiments' condition in Micro-blogging Bi-Followed Network. Exp. 5 executes the community detection in the pure Micro-blogging Bi-Followed Network; the relation strength of User Interactive Frequency is added in Exp. 6; Exp. 7 combines both the relation strength of User Attribute Similarity and User Interest Similarity; and Exp. 8 includes all the strength using the stated multi-dimensional network algorithm.

The results are shown in Table **5**. Obviously, Exp. 5 got the worst community detection quality in this group of ex-

Table 5. Community detection results.

Indicator	Exp. 5	Exp. 6	Exp. 7	Exp. 8
Qs	0.5427	**0.6727**	0.6100	0.6540
Frequency	2.5589	2.7697	2.6286	**2.7949**
ContentSu	0.0398	0.0407	**0.0532**	0.0517
UserSu	0.4329	**0.4365**	0.4351	0.4352
$\bar{\mu}$	0.8936	0.9799	0.9317	**0.9840**
Time(s)	93.316	88.348	87.200	87.794

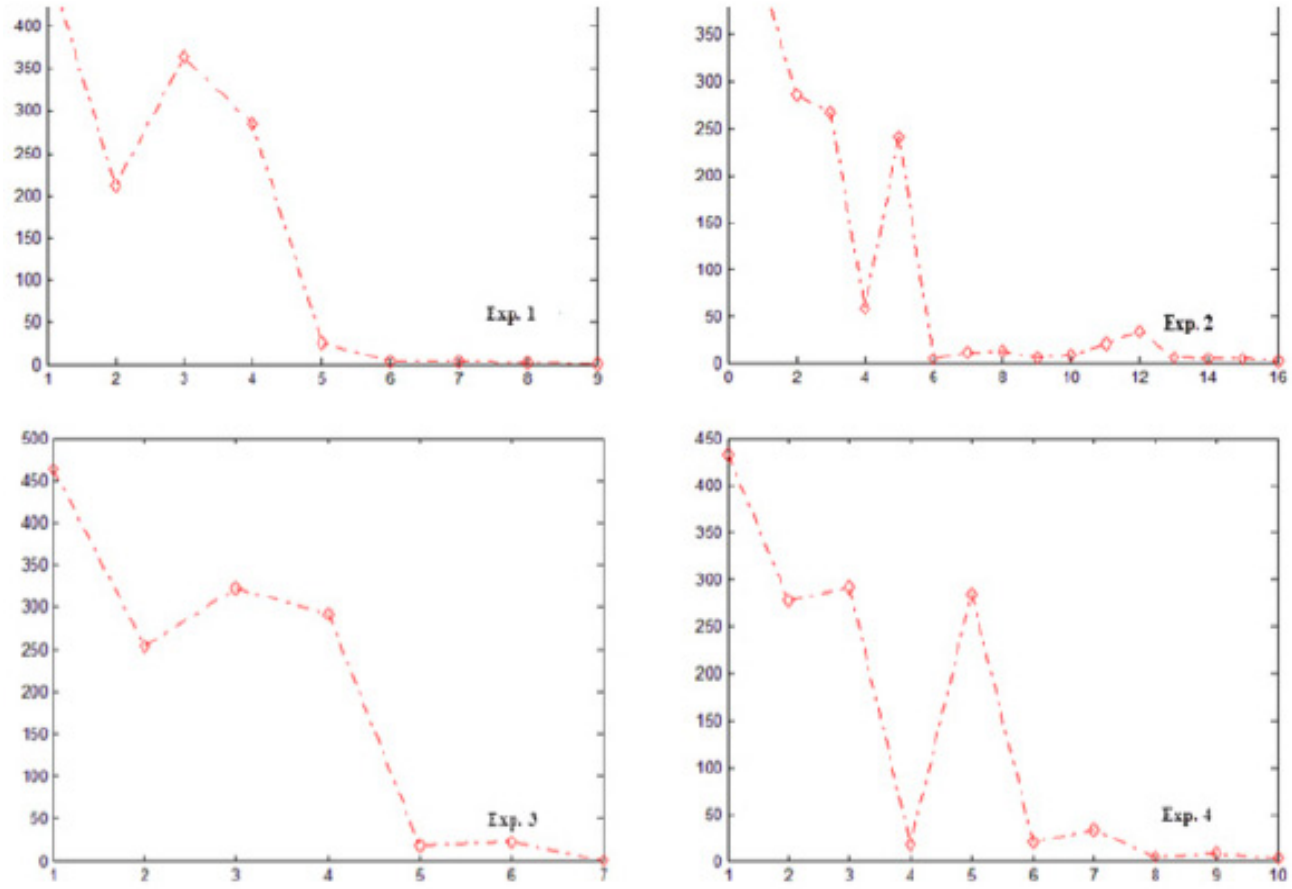

Fig. (2). Community detection results in micro-blogging single-followed network.

periments, but better than the counterpart in Micro-blogging Single-Followed Network. And it indicates better results in Micro-blogging Bi-Followed Network than Micro-blogging Single-Followed Network with the same strength conditions. Exp. 6 and Exp. 8 have the better value in *Qs*, *Frequency* and *UserSu*. And Exp. 6 values highest in *UserSu*. All of these means User Interactive Frequency is also a key strength in Micro-blogging Bi-Followed Network.

Furthermore, Exp. 8 got the highest score in $\bar{\mu}$, which indicates the comprehensive community detection quality of the result. It illustrates that in Micro-blogging Bi-Followed Network, more information will result in better community detection results.

Fig. (**2**) shows the number of division communities and the scale of each community in Micro-blogging Single-Followed Network. X-axis indicates the index of the divided communities, while y-axis indicates the scale of the communities. Exp. 1 divided the primary network into 9 communities, while Exp. 2 has 16 communities, Exp. 3 has 7 communities and Exp. 4 has 10 communities. It reveals that the di-

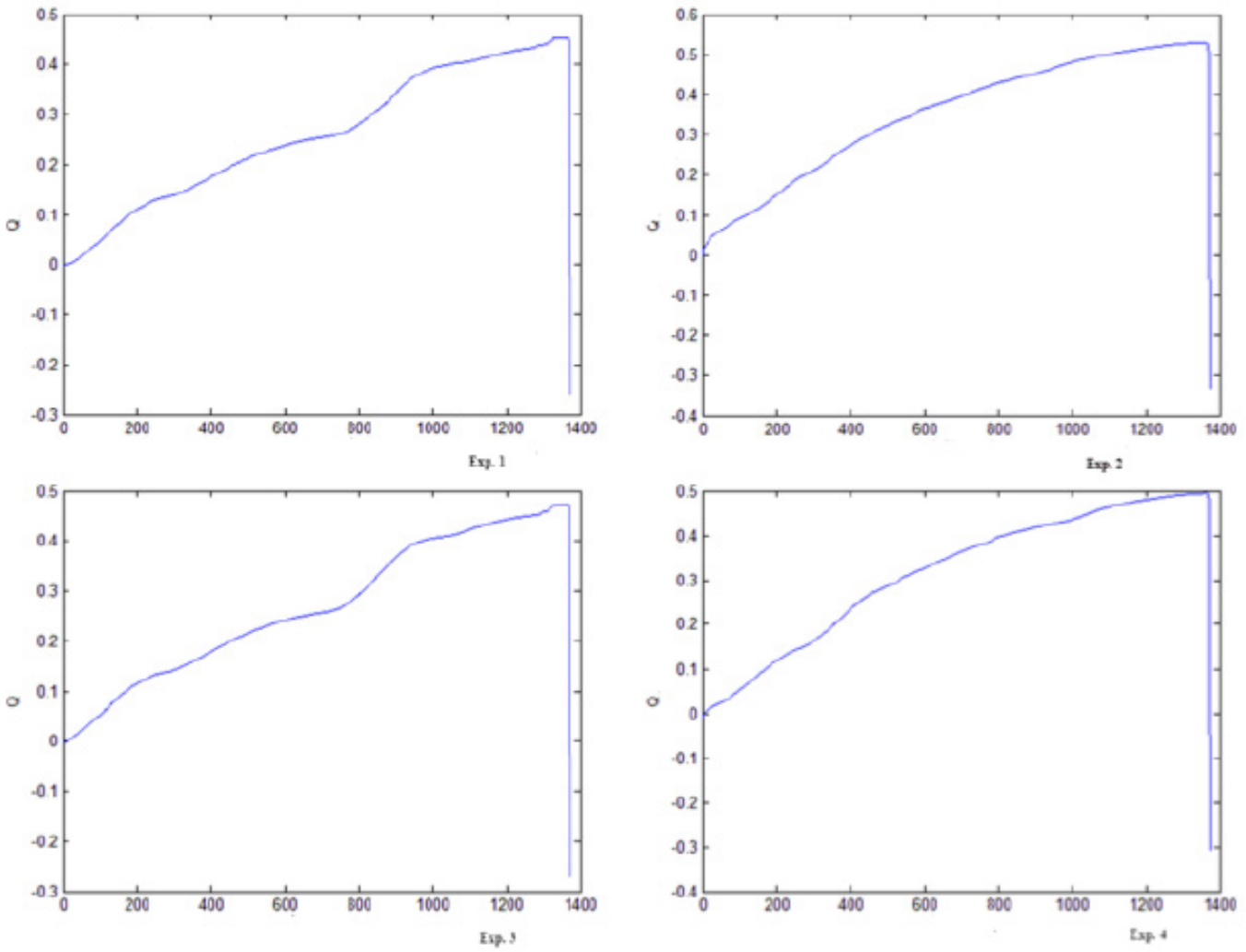

Fig. (3). Community detection results in micro-blogging single-followed network.

vision results turned to be balance in Exp. 2, 3, and 4, and User Interactive Frequency is a strict division factor, which increases the number of division communities. The introduction of User Attributes/Interest Similarity makes the closer relationship in the same community, and reduces the number of the division communities. Exp. 4 balances the number and the scale of the communities.

Fig. (**3**) shows the community detection results in Micro-blogging Single-Followed Network. X-axis shows the iterations (the number of the users in the experiment); while y-axis shows the Modality *Qs*.

Fig. (**4**) shows the experiments' results in Micro-blogging Bi-Followed Network. And it's a lot different from the results in Micro-blogging Single-Followed Network. With the same strength of edges, experiments in Micro-blogging Bi-Followed Network got much more communities, which is around 2 to 3 times, than that in Micro-blogging Single-Followed Network. It testifies to the truth that a stricter formed network makes the longer distance among users, and the weak ties are neglected; finally it bring forth more division communities.

While Exp.6 has 23, Exp. 7 has 29 and Exp. 8 has 21. Easily seen from the Fig. (**4**), Exp. 5 and Exp. 7 don't obtain balance classification results, while the results in Exp. 6 and Exp. 8 seem much better. Different from Micro-blogging Single-Followed Network, User Interactive Frequency aggregates the users in Micro-blogging Bi-Followed Network, and the User Attribute/Interest Similarity makes an increase in the number of division communities.

Fig. (**5**) shows the community detection results in Micro-blogging Bi-Followed Network. X-axis shows the iterations (the number of the users in the experiment); while y-axis shows the Modality *Qs*.

CONCLUSION

Existing community detection methods are mostly based on the analysis of the links between nodes ignoring the rich information of new era social media, the others clustering methods sometime ignore the network structure which is the foundation of social media. Aiming at the existed problems on community detection of Micro-blogging, this paper proposed a community detection algorithm based on multi-dimensional weighted network.

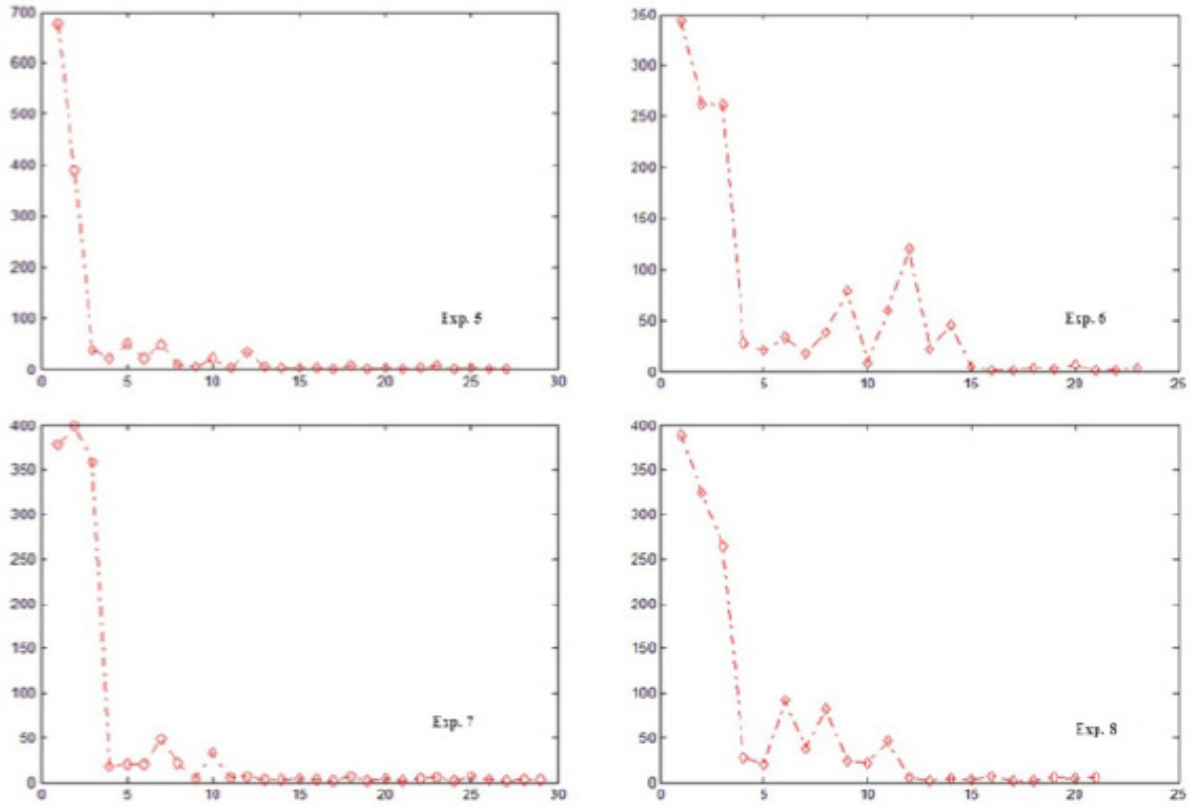

Fig. (4). Community detection results in micro-blogging bi-followed network.

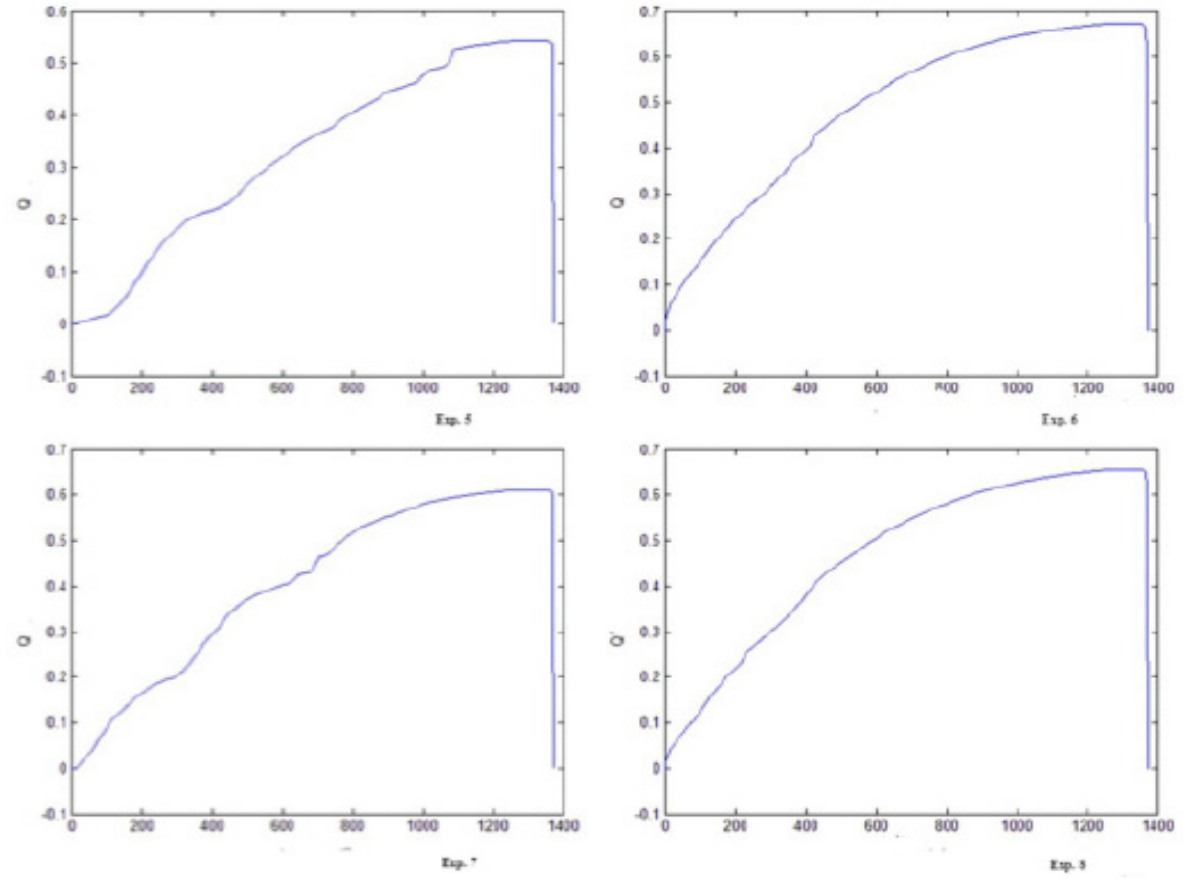

Fig. (5). Community detection results in micro-blogging bi-followed network.

Firstly, two basic network topologies of Chinese Micro-blogging are introduced, which are called Micro-blogging Bi-Followed Network and Micro-blogging Single-Followed Network. Secondly, the weighting methods are stated, and three edge betweennesses are defined: User Interactive Frequency, User Interest Similarity, and User Attributes Similarity. Combing the edge betweennesses with the basic network forms a multi-dimensional network. Then, an algorithm, which converts the multi-dimensional network into a single-dimensional network, is proposed.

Also, this paper constructed an comprehensive evaluation indicator $\bar{\mu}$, which is the mean value of the moudulity *Qs*, user interactive frequency within the community *Frequency,* user interest similarity within the community *ContentSu,* user attribute similarity within the community *UserSu*.

The Experimental results show that the classification results are better when extra information is used. And for Chinese Micro-blogging platform, the User Interactive Frequency plays a much more important role in community detection than user Interest Similarity and User Attribute Similarity.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The research was supported by Beijing Higher Education Young Elite Teacher Project (No. YETP1660), the Natural Science Foundation of Beijing under grant no. 4132067 and the Natural Science Foundation of China under grant no. 71271211.

REFERENCES

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

[2] S. White, and P. Smyth, "A spectral clustering approach to finding communities in graph," *Proceedings of the Fifth SIAM International Conference on Data Mining*, pp. 274–285. 2005.

[3] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", *Nature,* vol. 435, pp. 814–818, 2005.

[4] X. He, H. Zha, C. H. Q. Ding, and H. D. Simon, "Web document clustering using hyperlink structures," *Computational Statistics and Data Analysis,* vol. 41, no. 1, pp. 19–45, 2002.

[5] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, "Probabilistic author-topic models for information discovery," *The proceeding of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 306–315, 2004.

[6] D. Zhou, I. Councill, H. Zha, and C. Giles, "Discovering temporal communities from social network documents," *Proceedings of the 7th IEEE International Conference on Data Mining*, pp. 745–750, 2007.

[7] F. Yan, M. Zhang, Y. W. Tan, J. Tang, and Z. H. Deng, "Community discovery based on actors'interests and social network structure," *Journal of Computer Research and Development*, vol. 47, pp. 357-362, 2010.

[8] A. Java, X. D. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56-65, 2007.

[9] Z. F. Zhang, Q. D. Li, D. Zeng, and H. Gao, "User community discovery from multi-relational networks," *Decision Support Systems*, vol. 54, no. 2, pp. 870-879, 2013.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014