SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2017, Vol. 60 092105:1–092105:14
doi: 10.1007/s11432-015-0899-6

Detecting micro-blog user interest communities through the integration of explicit user relationship and implicit topic relations

Yu QIN[1,2], Zhengtao YU[1,2]*, Yanbing WANG[1,2], Shengxiang GAO[1,2] & Linbin SHI[1,2]

[1] *Institute of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650500*, China;*
[2] *Key Laboratory of Intelligent Information Processing, Kunming University of Science and Technology, Kunming* 650500*, China*

Received April 19, 2016; accepted July 4, 2016; published online April 27, 2017

Abstract In order to effectively utilize the explicit user relationship and implicit topic relations for the detection of micro-blog user interest communities, a micro-blog user interest community (MUIC) detection approach is proposed. First, through the analysis of the follow relationship between users, we have defined three types of such relationships to construct the user follow-ship network. Second, taking the semantic correlation between user tags into account, we construct the user interest feature vectors based on the concept of feature mapping to build a user tag based interest relationship network. Third, user behaviors, such as reposting, commenting, replying, and receiving comments from others, are able to provide certain guidance for the extraction of micro-blog topics. Hence, we propose to integrate the four mentioned user behaviors that are considered to provide guidance information for the traditional latent Dirichlet allocation (LDA) model. Thereby, in addition to the construction of a topic-based interest relationship network, a guided topic model can be built to extract the topics in which the user is interested. Finally, with the integration of the afore-mentioned three types of relationship network, a micro-blog user interest relationship network can be created. Meanwhile, we propose a MUIC detection algorithm based on the contribution of the neighboring nodes. The experiment result proves the effectiveness of our approach in detecting MUICs.

Keywords feature mapping, implicit topic, guided topic model, contribution of the neighboring nodes

Citation Qin Y, Yu Z T, Wang Y B, et al. Detecting micro-blog user interest communities through the integration of explicit user relationship and implicit topic relations. Sci China Inf Sci, 2017, 60(9): 092105, doi: 10.1007/s11432-015-0899-6

1 Introduction

With the continuous development of the social network, more and more participants are becoming involved in micro-blogging. With such a vast user base, the social network has become a popular subject of research by several specialists and scholars with respect to analysis methods of social network data [1–3] and methods to accurately understand the user interest, detect user interest communities, and quickly and efficiently detect those users with similar interests. In terms of web community detection, scholars

* Corresponding author (email: ztyu@hotmail.com)

Table 1 Follow relationships between micro-blog users

Definition of follow relationships	Description	Relationship strength (Se)
Follow each other	A follows B and B follows A	1.0
Unidirectional	A follows B or B follows A	0.7
Strangers	Without follow relationship	0.3

at home and abroad have conducted numerous studies in recent years. Mrinmays et al. [4] proposed a generative model to detect communities based on the topic discussions, interaction types, and social connections among people. Sun and Lin [5] proposed a probabilistic generative model to detect the latent topical communities among the users through the capture of user tagging behaviors and interests. Zhao et al. [6] proposed a topic oriented community detection approach combining social object clustering with link analysis. The experiments on real data sets had shown that their approach was able to identify more meaningful communities. Jin et al. [7] proposed a Markov dynamics based on the unfold and extract overlapping communities (UEOC) algorithm for the detection of overlapping communities in complex networks, and the experimental result had shown that UEOC was highly effective. Emanuele et al. [8] proposed an information dynamics algorithm to detect communities in networks, based on information diffusion. This algorithm included a non-linear processing phase. Wu et al. [9] discussed how profile information could be used to improve community detection in online social networks. In their study, they extended two popular community detection algorithms on which the profile information of users was utilized to partition online social networks. Ruan et al. [10] proposed an approach to detect the communities through a combination of content with link information in graph structures. The experiments had shown that good results had been achieved through their approach. Li and Pang [11] proposed a unified community detection algorithm for a complex network. The experimental results showed that their model was able to find the community structure successfully without using the network structure type. Wu and Zou [12] proposed an incremental community detection method for social tagging systems based on locality-sensitive hashing, the experimental results indicated that their method was able to detect communities more efficiently and effectively. Xin et al. [13] proposed a clustering algorithm for community detection based on the link-field-topic model to solve the issue of presetting the number of communities. Good results have been achieved in community detection through all the afore-mentioned approaches. In addition to the explicit user relationship, such as the user-follow relationship and user tag, this paper takes the implicit topic relations including user behaviors and micro-blog topics into comprehensive consideration to construct a micro-blog user interest relationship network. Additionally, by combining the concept of local information with the characteristics of the micro-blog user interest relationship network, we propose a MUIC detection method based on the contribution of neighboring nodes.

2 Explicit user relationship

2.1 Constructing the user-follow relationship network

Mutual follow relationships are commonly found among the micro-blog users. As a follow relationship is able to reflect the user's interest tendency to some extent, we shall define three types of follow relationships as shown in Table 1 through the analysis of the follow relationship between micro-blog users.

In Table 1, the relationship strength depends on a series of tests and corresponds to each type of follow relationship, declining in turn. The mutual follow relationship between users indicates that the users have an extremely similar interest tendency. In this case, the relationship strength can be set as 1. The unidirectional follow relationship between users indicates that the users have only a similar interest tendency. In this case, the relationship strength can be set as 0.7. Where the users are strangers, there is low possibility for the users to have a similar interest tendency. Hence, we might define such relationship strength as 0.3. The user follow relationship network is indicated in Figure 1.

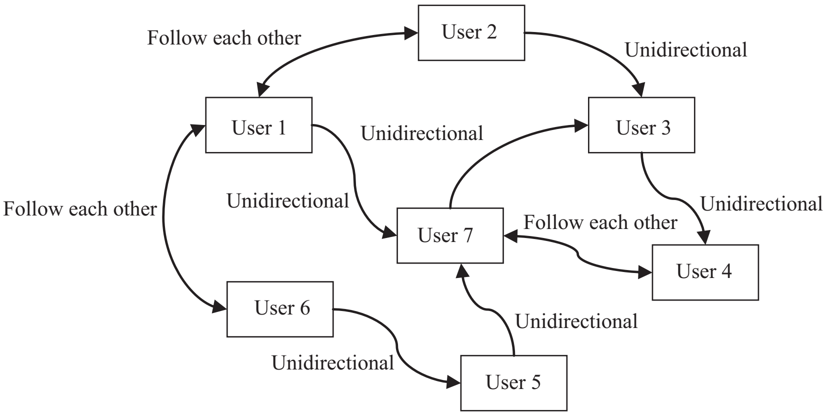

Figure 1 Micro-blog user-follow relationship network.

2.2 Constructing the user tag-based interest relationship network

When a user registers for a micro-blog account, the system will guide the user to fill in a personalized tag. The analysis of the personalized user tag information reveals that the tag is able to reflect the user's interest tendency to some extent. Utilizing this information, it is feasible to construct a user tag-based interest relationship network by constructing a user interest feature vector for every user. Additionally, the vector similarity can be used to represent the strength of the user interest relationship.

2.2.1 *Feature selection*

Every user and followee tag can be obtained through the Sina (Chinese online media company) micro-blog API. However, because user tag is created in a semi-guided mode, the personalized tag completed by the user will show strong randomness. Under such circumstances, it would be difficult to analyze and process this type of tag. Therefore, in order to obtain a tag set containing all user tags that meet the requirements, we propose to remove all personalized tags attached with special symbols or containing English words. Then, with the help of the self-developed statistical tool for word frequency of user tags on the Sina micro-blog, the frequency of the occurrence of all tags can be calculated. Once all tags have been ranked in descending order according to frequency, we will be able to choose those tags with top ranking as the feature dimension of user interest feature vectors by setting a threshold.

2.2.2 *Feature mapping-based representation of user feature vectors*

Due to the limitation of the dimensions, it is not possible for the feature dimension chosen to represent the user interests to cover all the tag words. Therefore, some information of guiding significance for the construction of user interest feature vectors might be lost. For example, the "travel" feature dimension can be found in the feature space derived through statistics. However, for some users, the "travel" tag might not be found in the tag set. In this case, the characteristic value of the "travel" feature dimension is zero for these users. Instead, feature analysis reveals that there might be a tag very similar to "travel" in the user's tag set, such as the "tour" tag. Under these circumstances, it is feasible to take the calculation of word similarity between "travel" and "tour" into account. Then, map the information of the "tour" tag onto the feature dimension of the "travel" tag through a mapping method to get closer to the user's real interest. Therefore, we introduce the concept of feature mapping to map the user tag onto every feature dimension according to the semantic similarity between the user tag and tag of the feature dimension. With this method, the corresponding characteristic value of every feature dimension can be calculated. During the process of feature mapping, as word similarity cannot normally be directly calculated between long tags, we utilize the ICTCLAS (Institute of computing technology, Chinese lexical analysis system) to segment the words. Hence, tags can be represented by a word set to calculate the average semantic similarity between the user tag and tag of the feature dimension. The specific approach is provided as follows. In order to process the tags in a consistent manner, we define all tags uniformly. The tags will be represented by a word set $l_u = \{\mathrm{wu}_1, \mathrm{wu}_2, \ldots, \mathrm{wu}_m\}$, where m is the number of the words contained in the tag, regardless of length. Similarly, the tag of every feature dimension can be represented by a word set $l_d = \{\mathrm{wd}_1, \mathrm{wd}_2, \ldots, \mathrm{wd}_n\}$, where n is the number of words contained in the tag of the feature dimension. Assume that the number of tags for every user is represented by X and frequency of the occurrence of a tag is represented by x. Then, f_{ul}, the initial characteristic value of every tag, can be expressed by

$$f_{\mathrm{ul}} = \frac{x}{X}. \tag{1}$$

Use $\mathrm{Sl}(l_u, l_d)$ to represent the semantic similarity between the user tag and tag of the feature dimension, where l_u represents the user tag and l_d refers to the tag of the feature dimension. The computational formula can be indicated as

$$\mathrm{Sl}(l_u, l_d) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Sim}(\mathrm{wu}_i, \mathrm{wd}_j)}{mn}, \tag{2}$$

where $\mathrm{Sim}(\mathrm{wu}_i, \mathrm{wd}_j)$ represents the average semantic similarity between user tag $\mathrm{wu}_i (i = 1, 2, \ldots, m)$ that contains m words and tag of feature dimension $\mathrm{wd}_j (j = 1, 2, 3, \ldots, n)$ containing n words. The word similarity can be calculated according to the HowNet-based word similarity computation method proposed by Liu [14]. Calculate the semantic similarity between each user tag and feature dimension tag, with the characteristic values yet to be determined. Then, choose a user tag that is most similar to a tag of the feature dimension and multiply the characteristic value of this tag by the maximum similarity to derive the result. The result can be considered as the characteristic value of this feature dimension. Thus, we obtain one of the feature dimensions. The above process is repeated to determine the characteristic value of every feature dimension. As a result, we can achieve feature mapping from the user tags to tags of the feature dimensions. The calculation of the characteristic value for each feature dimension is indicated as (3) in the feature mapping process:

$$T(l_d) = f_{\mathrm{ul}}((l_u)_a)\max\{\mathrm{Sl}((l_u)_a, l_d)\}, \ \ (a = 1, 2, \ldots, X), \tag{3}$$

where $\max\{\mathrm{Sl}((l_u)_a, l_d)\}$ represents the maximum similarity after the separate calculation of similarity between the X tags for a user and feature dimension l_d. $f_{\mathrm{ul}}((l_u)_a)$ is the characteristic value of the user tag when the maximum similarity between $(l_u)_a$ and l_d is derived. $T(l_d)$ is the characteristic value of feature dimension l_d.

2.2.3 *The user tag-based interest relationship network*

Based on the concept of feature mapping, all feature vectors representing user interest can be constructed for each user. In order to represent the strength of the interest relationship between different users, we use the cosine similarity calculation method to calculate the feature vector similarity between users. The computational formula is shown as

$$\mathrm{sim}\,(u_1, u_2) = \cos\,(\boldsymbol{u}_1, \boldsymbol{u}_2) = \frac{\boldsymbol{u}_1 \cdot \boldsymbol{u}_2}{\|\boldsymbol{u}_1\|\,\|\boldsymbol{u}_2\|}. \tag{4}$$

Taking the user as the node and feature vector similarity between the users as an edge, we construct a user tag-based interest relationship network. The thickness of the edge represents the degree of the interest similarity between the users. Figure 2 shows the user tag-based interest relationship network.

3 Implicit topic relations

3.1 Constructing the topic-based interest relationship network

Because a micro-blog published by a user always contains a topic in which the user is interested, it can be represented through the implicit topic relations between the micro-blogs. Therefore, on this account, we consider utilizing LDA [15] to extract the micro-blog topic for the representation of user

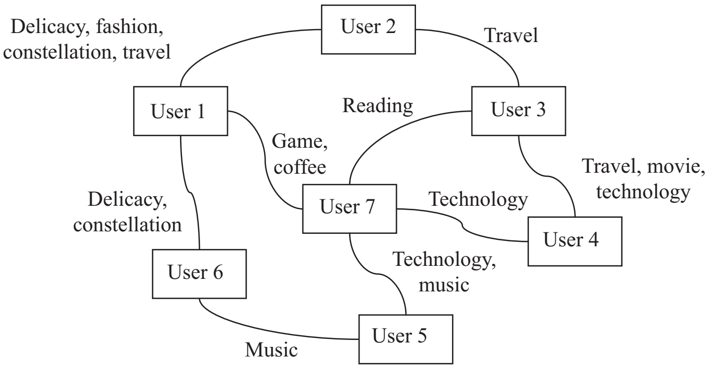

Figure 2 The user tag-based interest relationship network.

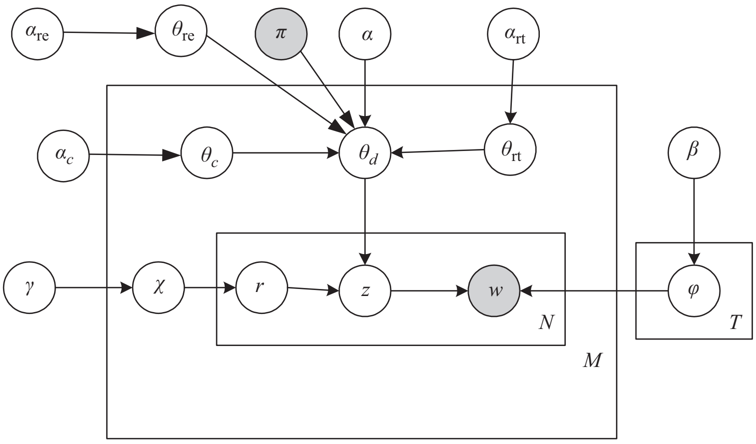

Figure 3 The Bayesian network diagram.

interest. During topic extraction, micro-blog user behaviors, such as reposting, commenting, replying, and receiving comments from others, will provide certain guidance for the extraction of implicit topics. Thus, we propose a guided LDA model for the extraction of user interest topics.

3.1.1 *Guided LDA-based micro-blog user interest modeling*

Based on the traditional LDA, we take user reposting, commenting, replying, and receiving comments from others as the guidance information for unified modeling. Then, integrated with the above guidance information, an LDA-based topic extraction model can be constructed. The Bayesian network diagram of this model is shown in Figure 3.

α, α_c, $\alpha_{\rm rt}$, and $\alpha_{\rm re}$ are the hyper-parameters of θ_d, θ_c, $\theta_{\rm rt}$, and $\theta_{\rm re}$, respectively, and separately represent the original micro-blog topic distribution, micro-blog comments, reposted micro-blog, and micro-blog comment replies, respectively. γ is the parameter for the extraction of χ distribution, while χ is the influence distribution of the relevant micro-blog sampled according to the Dirichlet distribution of parameter γ for all relevant commented micro-blogs. r is the influential micro-blog extracted from the χ distribution, and w represents the words contained in the micro-blog. N denotes the total of N words that have been contained. M indicates that there are a total of M micro-blog pieces. β is the hyper-parameter of φ, given as the word distribution under the topic. T indicates that there are a total of T topics. First, this model extracts φ, the relationship between the topic and words, from the Dirichlet distribution of parameter β. Then, the type of micro-blog will be determined according to the data source and data form when a piece of micro-blog is generated:

(1) If the micro-blog is original without any comment, set the value of π as 0, to represent that this micro-blog is original. At this point, it is necessary to obtain θ_d, given as the relationship between this micro-blog and selected topics from the Dirichlet distribution of parameter α.

(2) If the micro-blog contains a user comment, set the value of π as 1, to represent that this is a commented micro-blog. At this point, it is necessary to obtain θ_c, given as the relationship between the commented micro-blog and selected topics from the Dirichlet distribution of parameter α_c through sampling. Meanwhile, assign the value of θ_c to θ_d, given as the relationship between micro-blog d and the selected topics.

(3) If the micro-blog has been reposted, set the value of π as 2, to represent that this is a reposted micro-blog. At this point, it is necessary to obtain $\theta_{\rm rt}$, given as the relationship between the micro-blog that has been commented and selected topics from the Dirichlet distribution of parameter $\alpha_{\rm rt}$ through sampling. Moreover, assign the value of $\theta_{\rm rt}$ to θ_d, given as the relationship between micro-blog d and the selected topics.

(4) If the micro-blog user comment has received a reply, set the value of π as 3. This indicates that the topic distribution of this reply has been influenced by the original comment and micro-blog receiving the comment. Further, the weight of influence on the topic distribution of this reply from the original comment and micro-blog is different, when μ, the influencing parameter, is introduced to represent the influencing weight of the commented micro-blog that received a reply. The influencing weight of the commented micro-blog is then given as $1-\mu$. Here, perform an integrated calculation on the topic distribution of both the commented micro-blog that received a reply and original commented micro-blog to derive a mixed topic distribution. Next, assign the calculated result to θ_d, given as the relationship between micro-blog d and the selected topics.

Therefore, the computation formula for the topic distribution of the afore-mentioned micro-blog types can be expressed as

$$p\left(\theta|\alpha,\mu\right)=\begin{cases}\theta_d, & \pi=0,\ \alpha=\alpha,\\ \theta_c, & \pi=1,\ \alpha=\alpha_c,\\ \theta_{\rm rt}, & \pi=2,\ \alpha=\alpha_{\rm rt},\\ \mu\theta_{\rm re}+\left(1-\mu\right)\theta_c, & \pi=3,\ \alpha=\{\alpha_{\rm re},\ \alpha_c\}.\end{cases} \tag{5}$$

Additional comments received from other users on the original micro-blog will always influence, to some extent, the topic distribution of this micro-blog. Therefore, we shall include this original micro-blog and comments in data set S_d. An χ_d distribution obtained through the sampling of the Dirichlet distribution of parameter γ will be added to this data set as well. For every word contained in the micro-blog text, first extract an influential micro-blog r, which is included in data set S_d, constituted by the original micro-blog and comments, from the χ_d distribution. Then, obtain θ_γ given as the relationship between this influential micro-blog and selected topics through the sampling of the Dirichlet distribution of parameter γ_d, to extract the topic z based on such relationship. Next, extract a word from φ to fill out the corresponding space on the micro-blog. In this model, the computation formula for the probability distribution of θ is shown as

$$P(\theta|\alpha,\mu,\gamma_d)=xP(\theta|\alpha,\mu)+(1-x)P(\theta_\gamma|\gamma_d), \tag{6}$$

where x is a Boolean value, and can be set to 1 when the micro-blog has been reposted, has received a comment or a reply to the comment. Otherwise x should be set to 0, representing that this micro-blog is original.

3.1.2 *Model deduction and the extraction of user interest topic*

The joint probability distribution of the micro-blog, words, and topic can be formally expressed as

$$P(r,z,w|\phi,\theta,\chi)=\prod_{d\in D}\frac{\Delta(M_d^r+\gamma_d)}{\Delta(\gamma_d)}\prod_{d\in D}\frac{\Delta(N_d^z+\alpha)}{\Delta(\alpha)}\prod_{z\in T}\frac{\Delta(N_z^w+\beta)}{\Delta(\beta)}, \tag{7}$$

where M_d^r is the counting vector of the words that influence the micro-blog and have been contained in micro-blog d. N_d^z is the counting vector of the observable topics that have been influenced by micro-blog

d. N_z^w is the counting vector of the words contained in topic z. Decompose (7) and iterate the Gibbs sampling according to (8):

$$P(z_i=z',r_i=r'|z_{-i},r_{-i},w)=\frac{P(z,r,w)}{P(z_{-i},r_{-i},w_{-i})}\frac{1}{P(w_i|z_{-i},r_{-i},w_{-i})}. \tag{8}$$

Obtain a posterior distribution formula that can be indicated by

$$\begin{aligned}P(z_i=z',r_i=r'|z_{-i},r_{-i},w)&\propto\frac{P(z,r,w)}{P(z_{-i},r_{-i},w_{-i})}\\&=\frac{N_{r'z'}^{-i}+\alpha}{N_{r'}^{-i}+T\alpha}\frac{M_{dr'}^{-i}+\gamma_d(r')}{\sum_{r\in S_d}(M_{dr'}^{-i}+\gamma_d(r))}\frac{N_{z'w_i}^{-i}+\beta}{N_{z'}^{-i}+V\beta},\end{aligned} \tag{9}$$

where $N_{r'z'}^{-i}$ is the co-occurrence frequency of the influential micro-blog r' and topic z'. $N_{r'}^{-i}$ represents the co-occurrence frequency of the influential micro-blog r' and all the topics. $N_{z'w_i}^{-i}$ is the co-occurrence frequency of the word w_i and topic z'. $N_{z'}^{-i}$ stands for the co-occurrence frequency of topic z' and all the words. For different micro-blog types α and β will have different parameter values. Continue to iterate (7) with a sampling of all the topics until the sampling results become stable. Because the sampling of words and topics meets the requirement of multinomial distribution, the results of θ_d, θ_c, $\theta_{\rm rt}$, $\theta_{\rm re}$, φ_z, and χ can be separately expressed by (10)–(15), respectively, as indicated below:

$$\theta_d=\frac{N_{\rm dz}+\alpha}{N_d+T\alpha}, \tag{10}$$

$$\theta_c=\frac{N_{\rm cz}+\alpha_c}{N_c+T\alpha_c}, \tag{11}$$

$$\theta_{\rm rt}=\frac{N_{\rm rtz}+\alpha_{\rm rt}}{N_{\rm rt}+T\alpha_{\rm rt}}, \tag{12}$$

$$\theta_{\rm re}=\frac{N_{\rm rez}+\alpha_{\rm re}}{N_{\rm re}+T\alpha_{\rm re}}, \tag{13}$$

$$\varphi_z=\frac{N_{\rm zw}+\beta}{N_z+V\beta}, \tag{14}$$

$$\chi_d\left(r\right)=\frac{M_{\rm dr}+\gamma_d\left(r\right)}{\sum_{r\in s_d}\left(M_{\rm dr}+\gamma_d\left(r\right)\right)}. \tag{15}$$

Obtain the micro-blog topic distribution through the Gibbs sampling. For every user, it is possible to obtain the user-topic distribution when the probability of each topic has been summed up, and the result has been divided by the number of the user's micro-blogs. Meanwhile, the user-topic feature vectors can be obtained. Additionally, for every topic, the user-topic probability can be calculated through

$$P_u(z_i)=\frac{\sum_{i=1}^{N}P(z_i)}{N}. \tag{16}$$

3.1.3 *The topic-based interest relationship network*

Take the user as the node and interest similarity as the edge, and a micro-blog topic-based interest relationship network can be built. The thickness of the edge represents the degree of interest similarity between the users. Figure 4 shows the micro-blog topic-based interest relationship network.

4 Our approach

4.1 Network convergence

Use S_e, with the value defined in Table 1 to represent the strength of interest relationship in the user follow-ship network. Use $S_{\rm hL}$ to represent the strength of interest relationship in the user tag-based

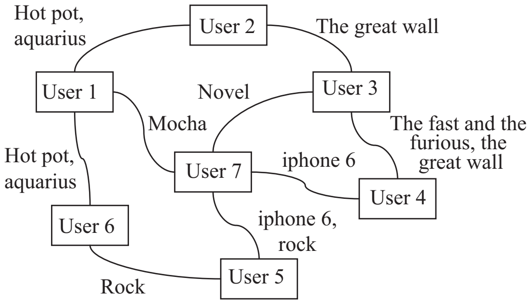

Figure 4 The micro-blog topic-based interest relationship network.

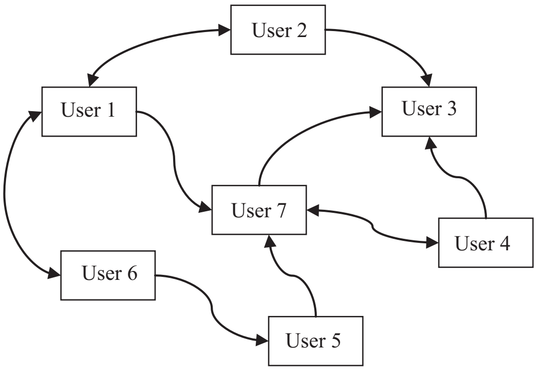

Figure 5 The user interest relationship network.

interest relationship network. Further, use S_{hM} to represent the strength of interest relationship in the micro-blog topic-based interest relationship network. Then, the total strength of the interest relationship between users can be formally defined through

$$S_t = S_e(\lambda_{\mathrm{hL}} S_{\mathrm{hL}} + \lambda_{\mathrm{hM}} S_{\mathrm{hM}}), \tag{17}$$

where λ_{hL} is the influential parameter for the strength of interest relationship based on the user tag, and λ_{hM} is the influential parameter for the strength of interest relationship based on the micro-blog topic. Through the analysis of the micro-blog tag and content, it is revealed that the user interest reflected by the tag has a certain hysteresis quality. However, the user interest reflected by the micro-blog content is more close to reality. Accordingly, the values of both parameters are separately set to be $\lambda_{\mathrm{hL}} = 0.3$ and $\lambda_{\mathrm{hM}} = 0.7$ to calculate S_t; the total strength of interest relationship between the users. Meanwhile, by setting a threshold value, the edges with the total strength of interest relationship lower than the threshold are deleted. As shown in Figure 5, a user interest relationship network can be finally constructed without consideration of the weight.

4.2 Detecting micro-blog user interest communities

In a practical micro-blog environment, there are many users exerting great influence. We can observe this when a user is followed by numerous users, or a post published by a user on his/her micro-blog is reposted many times. Such following or reposting behavior, in itself, reflects the user's interest tendency. Therefore, all such users might be within the same community of interest. Based on this, a user interest community can be built based on these influential users when numerous users closely related to them are continuously attracted into the community. According to the above analysis, we propose a MUIC detection method based on the contribution of the neighboring nodes.

4.2.1 *The main idea of the method*

When those users exerting great influence on the micro-blog are mapped into the user interest relationship network, they will be represented as nodes with a high indegree. Users showing interest tendencies similar to the influential users will be mapped into the user interest relationship network and represented as neighboring nodes to those with high indegree. The degree of user interest tendency determines if a user can be covered in an interest community, as it can be represented by the contribution of neighboring nodes. We provide a community detection process through our MUIC detection method as follows. Take the node with the highest indegree in the network as the initial community. Then, calculate the node contribution of all neighboring nodes in this initial community. Include all nodes with a node contribution greater than zero and those with the greatest contribution in this community. The community is deemed saturated if the contribution is lower than 0. Repeat the above process until all communities of interest are detected. Through the above community detection process, all detected communities will comply with the characteristics of nodes; densely connected within the community, yet sparsely connected in different communities. Moreover, integrated with the concept of local information, this algorithm has low complexity as it only considers the neighboring nodes of the initial community instead of the whole network. However, the detected communities might overlap through this method; as such, it is necessary to merge those communities with an overlapping degree larger than the threshold after the community overlapping degree is calculated.

4.2.2 *Node contribution*

Node contribution represents the degree of contribution made by the nodes to the community. The greater the node contribution, the more necessary the nodes will be covered in the community. The node contribution can be formally represented by

$$f_G^A = f_{G+\{A\}} - f_{G-\{A\}}, \tag{18}$$

where $f_{G+\{A\}}$ is the fitness of community G when node A is subordinate to G. $f_{G-\{A\}}$ is the fitness of G when node A is not covered in G. When $f_G^A > 0$, it denotes that node A makes a contribution to G. f_G, the fitness of G, can be formally represented by

$$f_G = \frac{k_{\rm in}^G}{(k_{\rm in}^G + k_{\rm out}^G)^\alpha}, \tag{19}$$

where $k_{\rm in}^G$ is twice to the sum of the edges when all of their endpoints are within the community, and $k_{\rm out}^G$ is the number of the edges when only one endpoint is located within the community. Use parameter α to control the scale of the community. The study of a practical network conducted by Lancichinetti et al. [16] revealed that when $\alpha < 0.5$, the whole network could be considered a community. Further, when $\alpha > 2$, each node in the network can be considered a separate community. For a large portion of the network, the value of α is in close proximity to 1.

In order to make the algorithm comply with characteristics of the micro-blog user interest relationship network, we revise f_G, the fitness of G. The revised f_G can be formally represented by

$$f_G = \frac{mk_{\rm d\_in}^G + nk_{\rm s\_in}^G}{(mk_{\rm d\_in}^G + nk_{\rm s\_in}^G + mk_{\rm d\_out}^G + nk_{\rm out\_in}^G + tk_{\rm in\_out}^G)^\alpha}, \tag{20}$$

where $k_{\rm d\_in}^G$ represents the value of all bidirectional nodes within G with a numerical value of twice the number of such node pairs. $k_{\rm s\_in}^G$ represents the value of nodes with a unidirectional follow relationship within G with a numerical value equal to the number of edges between such nodes. $k_{\rm d\_out}^G$ represents all the values of nodes that are connected to each other within and outside G with a numerical value equal to the number of such node pairs. $k_{\rm out\_in}^G$ refers to all the values of nodes that are outside, yet, connect inside to the nodes within G with a numerical value equal to the number of edges between such nodes. Finally, $k_{\rm in\_out}^G$ represents all the values of nodes that are inside G however, connect to nodes outside the community with the numerical value equal to the number of the edges between such nodes. Set the value of parameter α to 1. The parameters of m, n, and t are the contribution coefficients introduced for the different subordinate tendencies of the community. m represents the contribution coefficient of the nodes that are connected to each other, n refers to the contribution coefficient of the nodes that connect to the nodes within the community. t represents the contribution coefficient of the nodes that are within the community and connect to the nodes in other communities. In order to normalize the contribution coefficients, we define the contribution coefficient vector as $\{\beta_1^i, \beta_2^i, ..., \beta_k^i\}$, where β_k^i is the degree of contribution made by node i, when it is subordinate to G. It conforms to the requirement that the sum of coefficients for contributions made by a random node in G to such G is equal to 1. This can be formally represented by

$$\sum_{k=1}^{n} \beta_k^i = 1, \ \ 0 \leqslant \beta \leqslant 1, \tag{21}$$

where i is an arbitrary node in the network, and n is the number of contribution coefficients.

Table 2 Sina micro-blog data set

The number of users	The number of user tags	The number of micro-blogs
1000	8947	50000

4.2.3 *Community overlapping degree*

Community overlapping degree represents the overlapping degree between the communities. The greater the community overlapping degree, the larger the overlapped area between the communities. The community overlapping degree can be formally expressed by

$$\mathrm{overlap}(G_i, G_j) = \frac{|G_i \cap G_j|}{|\min(G_i, G_j)|}, \tag{22}$$

where $\min(G_i, G_j)$ represents the number of nodes in community G_i or G_j, whichever has the lowest number of nodes.

4.2.4 *The algorithm process*

(1) Calculate P, the value of all nodes in the network. Then, successively store the nodes ranked in descending order according to the value of P into queue Q.

(2) De-queue the top nodes from Q and take the front node N as the initial G.

(3) Calculate f_G^A, the node contribution made by all neighboring nodes in G. If the maximum node contribution made by any neighboring node is $\max(f_G^A) > 0$, then include this node with the greatest contribution into G and remove it from Q. Repeat this process. If $\max(f_G^A) < 0$, it denotes that the node contribution made by every neighboring node is less than 0. In this case, it indicates that G is saturated and all members have been selected into G. Skip to step (4) to detect other communities in the network.

(4) If Q is not empty, return to step (2). However, if Q is empty, it denotes that all communities in the network have been detected. At this point, skip ahead to step (5).

(5) Calculate the overlapping degree of all detected communities in pairs, and merge those communities whose overlapping degree is higher than the threshold value.

5 Experiments

5.1 Experimental data set

We manually selected 1000 celebrity users. Each of these users exerted a great influence on the Sina micro-blog. The crawling of the users' micro-blog information covers the follower list, user custom tags, followee tags, and posts published by the users on their respective micro-blogs. The basic information about the crawled experimental data set is shown in Table 2.

5.2 Pre-processing of experimental data set

5.2.1 *Pre-processing of user tag*

Construct a tag set consisting of user personal tags and followee tags for each user. Remove those tags attached with special symbols or containing English words to build a user interest feature vector for each user. Next, combine the tag sets of all users to build a tag library in order for word frequency to be calculated to determine the feature space.

5.2.2 *Pre-processing of micro-blog content*

(1) Symbol of @. Followed by nickname of a user, this symbol represents that this is a guided micro-blog. We choose to delete the content between two @ symbols.

(2) Symbol of //. This symbol can be primarily found in a reposted micro-blog to connect all attached micro-blogs reposted by several users. While we are trying to save such micro-blogs, the posts must be saved separately.

(3) Symbol of #. This represents the topic discussed by the user on the micro-blog with the structure indicated by # content of topic #. The content is always expressed through hyperlinks. Targeting such a structure, we choose to remove the content between the two # symbols.

(4) Hyperlink in text. In some micro-blogs, there might be only one or two words at the time they are published by users. For such micro-blogs, we choose to utilize web crawlers to save the webpage to which the URL corresponds, and then make an auxiliary analysis on the micro-blog.

5.2.3 *Evaluation index for community division*

Modularity function [17] is an evaluation criterion that has been widely recognized by scholars for detection of non-overlapping communities. Moreover, some of the community detection algorithms have been directly designed from the perspective of optimizing modularity function. However, the overlapping community detection methods are directly based on the modularity function. Therefore, some scholars have expanded the modularity function such that it can be used to detect communities by evaluating the overlapping community detection algorithms. For example, Sales-Pardo et al. [18] proposed EQ, the expanded modularity function and formally defined it by

$$\mathrm{EQ} = \frac{1}{2m}\sum\nolimits_{ij}\frac{1}{o_i o_j}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i c_j), \tag{23}$$

where m is the number of edges in the network, $2m$ is the total node degree in the network, and o_i refers to the number of communities that contain node i. The value of A_{ij} is of Boolean type. It may be set as 1, if there are edges available to connect node i and j, otherwise the value is 0. k_i is the degree of node i. The value of function $\delta(c_i c_j)$ is set to 1, if community c_i and c_j are the same community, otherwise the value is 0. Eq. (23) reveals that when the endpoints of an edge have been included in additional communities, the edge will play a more significant role in modularity; otherwise it will have less influence. When the communities have been detected through our MUIC detection algorithm, they might overlap during the detection of user interest communities. For this reason, we shall use EQ as the evaluation index for community division.

5.3 Experimental settings and result analysis

Experiment 1. Contrast experiment on the interest community division modularity based on different algorithms.

In order to verify the effectiveness of the MUIC detection method proposed in this paper, we chose to conduct a contrast experiment based on the clique percolation method (CPM) [19], community overlap propagation algorithm (COPRA) [20], and our approach. With the adoption of our approach, the parameters in (20) are set to be $m = 0.5$, $n = t = 0.25$, and $\alpha = 1$ in the experimental process. These three methods are separately used to detect the MUICs in the micro-blog user interest relationship network. The community detection results are shown in Figure 6.

Figure 6 indicates that for the detection task of the MUICs, our approach is slightly superior to the CPM and COPRA from the perspective of community division modularity. Such an experiment result can be easily explained in theory: In the community detection process, the CPM has to continuously detect the maximum sub-graph consisting of k-cliques. As these k-cliques are mutually connected in the network, this algorithm must be strictly defined. The COPRA will determine the community that the nodes are subordinate to according to the number of tags brought by the neighborhood set. However, it does not consider the situation that different neighboring nodes might give a different contribution to the community division. That is to say, the connection between the members in the community is much closer than the connection between the communities.

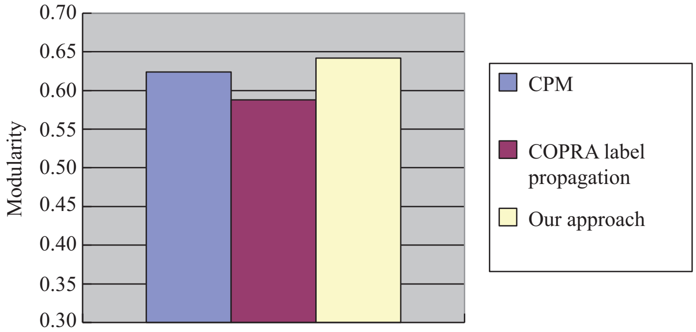

Figure 6 (Color online) Contrast experiment on the interest community division modularity based on different algorithms.

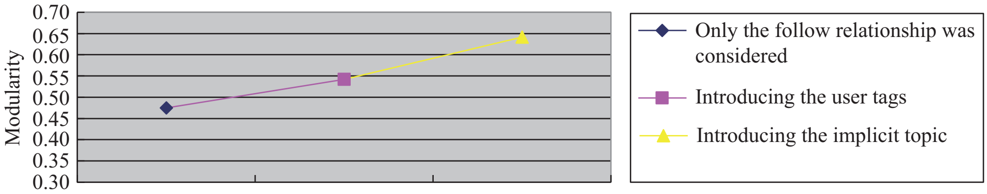

Figure 7 (Color online) Influence of explicit user relationship and implicit topic relations on the division of interest communities.

Experiment 2. Influence of explicit user relationship and implicit topic relations on the division of interest communities.

In order to verify the influence of the explicit user relationship and implicit topic relations on the division of interest communities, we designed a contrast experiment as follows. First, a follow-ship network was built considering only the follow relationship. Based on such a network, we divided the interest communities through the algorithm proposed in this paper and calculated the community division modularity. Next, in order to verify the influence of user tags on the interest community division, we introduced a user tag-based user interest relationship network based on the existing follow-ship network. Moreover, according to the algorithm proposed in this paper, we divided the interest communities to calculate the community division modularity. Finally, in order to verify the influence of the implicit topic relations on the division of interest communities, we introduced an implicit topic-based user interest relationship network based on the existing interest relationship network. According to the algorithm proposed in this paper, we divided the interest communities to calculate the community division modularity. The results of the contrast experiment in three phases are shown in Figure 7.

Figure 7 reveals that from the perspective of the interest community division effect, the modularity is maximized when both the explicit user relationship and implicit topic relations are taken into account during the community detection. However, the modularity is minimal when only the follow-ship is used to divide the communities in the explicit user relationship. The experiment result shows that a better result will be achieved in the division of MUICs when both the explicit user relationships and implicit topic relations are taken into account.

Experiment 3. Influence of tag feature mapping and guided LDA model on the division of interest community.

We have designed four contrast experiments with the experiment results shown in Figure 8. Figure 8 indicates that the modularity is maximized when communities have been divided according to the feature mapping-based and guided LDA-based methods proposed in this paper. However, the modularity is minimal when only the traditional LDA model is applied to divide the community without the introduction of feature mapping. The experiment result proves that both the tag feature mapping and guided LDA model that considers user behaviors will provide certain support to the detection of user interest communities.

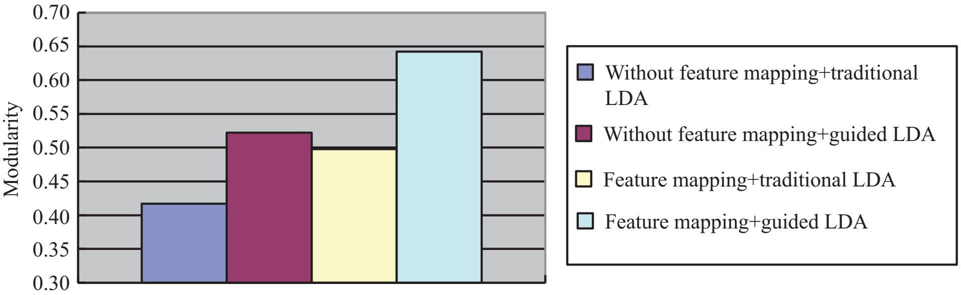

Figure 8 (Color online) Influence of tag feature mapping and guided LDA model on the division of interest community.

6 Conclusion

This paper analyzes the mutual follow relationship between users in a micro-blog environment. A user tag-based interest model is constructed according to the concept of feature mapping. Moreover, by taking the user behavior as supervisory information, we build a guided LDA-based topic extraction model to extract the micro-blog topics. Finally, a micro-blog user interest relationship network is constructed. We combine the concept of local information with the characteristics of the micro-blog user interest relationship network for the community detection algorithm. With this method, we propose a MUIC detection algorithm based on the contribution of neighboring nodes. The experiment result proves the effectiveness of our approach in the detection of MUICs.

Acknowledgements This work was supported by National Nature Science Foundation (Grant Nos. 61175068, 61472168), Key Project of Yunnan Nature Science Foundation (Grant No. 2013FA130), Key Special Project of Yunnan Ministry of Education, Ministry of Education of Returned Overseas Students to Start Research and Fund Projects, and Science and Technology Innovation Talents Fund Projects of Ministry of Science and Technology (Grant No. 2014HE001).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Howden C, Liu L, Li Z Y, et al. Virtual vignettesthe acquisition, analysis, and presentation of social network data. Sci China Inf Sci, 2014, 57: 032104
2 Yuan X H, Buckles B P, Yuan Z S, et al. Mining negative association rules. In: Proceedings of the 7th IEEE Symposium on Computers and Communications, Taormina, 2002. 623–628
3 Fei Z S, Ding H C, Xing C W, et al. Performance analysis for range expansion in heterogeneous networks. Sci China Inf Sci, 2014, 57: 082305
4 Sachan M, Contractor D, Faruquie T A, et al. Using content and interactions for discovering communities in social networks. In: Proceedings of the 21st International Conference on World Wide Web, Lyon, 2012. 331–340
5 Sun X L, Lin H F. Topical community detection from mining user tagging behavior and interests. J Am Soc Inf Sci Technol, 2013, 64: 321–333
6 Zhao Z Y, Feng S Z, Wang Q, et al. Topic oriented community detection through social objects and link analysis in social networks. Knowl Syst, 2012, 26: 164–173
7 Jin D, Yang B, Baquero C, et al. Markov random walk under constraint for discovering overlapping communities in complex networks. J Stat Mech: Theor Exp, 2011, 2011: 05031
8 Massaro E, Bagnoli F, Guazzini A, et al. Information dynamics algorithm for detecting communities in networks. Commun Nonlin Sci Numer Simul, 2012, 17: 4294–4303
9 Wu F Yeung K H. Incorporating profile information in community detection for online social networks. Phys A: Stat Mech Appl, 2014, 405: 226–234
10 Ruan Y Y, Fuhry D, Parthasarathy S. Efficient community detection in large networks using content and links. In: Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, 2013. 1089–1098
11 Li K, Pang Y. A unified community detection algorithm in complex network. Neurocomputing, 2013, 130: 36–43
12 Wu Z Y, Zou M. An incremental community detection method for social tagging systems using locality-sensitive hashing. Neur Netw, 2014, 58: 14–28
13 Xin Y, Yang J, Xie Z Q. A semantic overlapping community detection algorithm based on field sampling. Expert Syst Appl, 2015, 42: 366–375
14 Liu Q, Li S J. Word similarity computing based on How-net. Int J Comput Linguist Chin Lang Proc, 2002, 7: 59–76

15 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation. J Mach Learn Res, 2003, 3: 993–1022

16 Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks. New J Phys, 2009, 11: 033015

17 Newman M E J, Girvan M. Finding and evaluating community structure in networks. Phys Rev E, 2004, 69: 026113

18 Sales-Pardo M, Guimerà R, Moreira A A, et al. Extracting the hierarchical organization of complex systems. Proc Nation Acad Sci Unit Stat Am, 2007, 104: 15224–15229

19 Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435: 814–818

20 Raghavan U, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E, 2007, 76: 036106