Local Community Detection Using Social Relations and Topic Features in Social Networks

Chengcheng Xu[1], Huaping Zhang[1], Bingbing Lu[1] and Songze Wu[1]

[1] School of Computer Science, Beijing Institute of Technology, Beijing, China
xuchengcheng2015@nlpir.org, kevinzhang@bit.edu.cn,
lubingbing2015@nlpir.org, wusongze2015@nlpir.org

Abstract. Local community detection is an important research focus in social network analysis. Most existing methods share the intrinsic limitation of utilizing undirected and unweighted networks. In this paper, we propose a novel local community detection algorithm that fuses social relations and topic features in social networks. By defining a new social similarity, the proposed algorithm can effectively reveal the dynamic characteristics in social networks. In addition, the topic similarity is measured by Jensen–Shannon divergence, in which the topics are extracted from the user-generated content by topic models. Extensive experiments conducted on a real social network dataset demonstrate that our proposed algorithm outperforms methods based on social relations or topic features alone.

Keywords: Social Networks, Local Community Detection, Topic Model.

1 Introduction

Community detection has become an important research focus in the area of social network analysis, with the aim of exploring multiple subgraphs to closely connect nodes in the same subgraph and reduce the links between different subgraphs for applications, such as data mining, behaveior analysis, and knowledge discovery [1-2]. More specifically, global community detection, in which structural information about the entire network is required to divide it into a number of “internal” and “external” communities, has attracted substantial attention due to the promising empirical results that could be obtained. However, these approaches have some limitations: (1) from the perspective of network partitioning, it is difficult to obtain the needed structural information of the entire network for large and dynamic networks [3]; and (2) the computational complexity of a community detection algorithm is very high, even if the structure of a large-scale dynamic network can be obtained.

Compared with global community detection methods, local community detection methods can effectively mine community structures without prior network information. These methods can usually be understood as exploring community structures from given seed nodes. Therefore, for large-scale dynamic social networks, local community detection methods have obvious advantages in terms of computational

This work is sponsored by National Basic Research Program of China (973 Program, Grant No.: 2013CB32960601).

complexity and local characteristics. In practice, these local community detection methods are beneficial for public opinion monitoring, social network marketing, and personal friend recommendation.

However, most popular social networks are typical user-generated content (UGC) platforms, in which users often have both social relations and topic features. Social relations refer to the real interactions between users, including commenting, quoting (@), and retweeting. These are directed, weighted and dynamic link relationships. In addition, users in the same community often share common interests (i.e., topics), which can be extracted from UGC by topic models. Currently, the main research studies have explored the local community structure according to the undirected and unweighted relationships between different users. However, no mature algorithms to explore the local community structure by applying the above two features are available.

In this paper, we propose a new local community detection algorithm for social networks that is based on users' social relations and topic features. We define the social similarity between users and communities by utilizing their directed and weighted relationships. Then, the classical Jensen–Shannon divergence is used to calculate the topic similarity extracted from UGC with topic models. Finally, a novel algorithm based on the fusion of social relations and topic features is proposed for local community detection. We conduct extensive experiments on real social networks. The experimental results prove that our method performs effectively. In addition, comparative experiments involving different algorithm parameters also provide empirical guidance for practical applications.

The main contributions of this paper are as follows:

- This paper proposes a novel local community detection algorithm based on social relations and topic features in social networks.
- This paper defines the social similarity between the user and the community, taking advantage of directed and weighted information.
- The experimental results prove that our method performs effectively in a real social network dataset.

This paper is organized as follows. Section 2 reviews related works on local community detection and the topic model. Section 3 describes in detail the local community detection algorithm based on social relations and topic features. Section 4 presents the experimental results obtained from the real social network dataset. The entire paper is summarized in Section 5.

2 Related Work

The algorithm proposed in this paper is based on two user attributes in social networks: social relations and topic features. Thus, there are two lines of research related to our work, namely, local community detection and the topic model.

2.1 Local Community Detection Method

The purpose of local community detection is to explore the community network structure from a given seed node in a social network. In recent years, many researchers have proposed local methods for community detection. Some local community metrics have also been proposed, such as R [4], M [5], L [6] and LS [8]. Combining these methods can help us find correct community structures. However, the most serious drawback is that most algorithms are sensitive to the initial node, and as a result, a large number of outliers are introduced into the target community. Some papers attempt to set the initial node to the nearest core node for community detection [9-10], but some limitations remain to be overcome. For example, when the core user belongs to several communities, substantial noise (i.e., nodes that do not belong to the target community) will appear. [11] proposed a local optimization method using random seed nodes (i.e., the Local Fitness Method [*LFM*] algorithm) to solve the problem that some methods cannot find the hierarchy and overlap community structures. [12-14] used edge clustering for local community detection.

The similarity of these methods is that they use social relationships alone as edges to aggregate nodes. The difference of our approach is that it combines both social relations and topic features for local community detection, making the experimental results more accurate and useful.

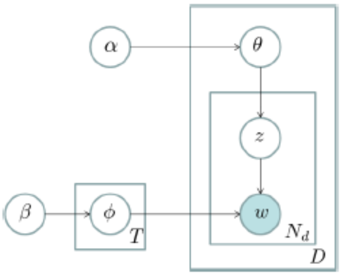

Fig. 1. Latent Dirichlet Allocation (LDA) bayesian network

2.2 Topic Model

LDA, a three-layer Bayesian generative probability topic model, assumes that documents are composed of a series of potential topics [15]. The topic is the abstract of all the words in the vocabulary. The main difference between documents is that they have different topic distributions. *LDA* treats documents as word frequency vectors using the bag-of-words model, which converts the text information into a digital representation. The process of document generation by the *LDA* model is shown in Figure 1.

$$p(\omega,z|\alpha,\beta)=p(\omega|\alpha,\beta)\,p(z|\alpha)=\int p(z|\theta)\,p(\theta|\alpha)d\theta\int p(\omega|z,\varphi)p(\varphi|\beta)d\varphi \tag{1}$$

Eq. (1) is the joint probability distribution for all words and topics in a document. The random variable θ is the topic vector distribution of the documents, which plays an important role in the calculation of topic similarity in section 3.2. When the correspondence between the user and the document is established, the topic similarity between users can be measured by Jensen–Shannon divergence. For dataset *D*, *LDA'*s topic extraction process is to maximize $p(D|\alpha,\beta)$ using *Gibbs Sampling*, *Variational Bayes* and *Expectation Propagation* [15-17].

3 Local Community Detection Using Social Relations and Topic Features

Let $G=(V,E)$ be the graphical representation of a directed social network, where *V* corresponds to the node set representing the users, and *E* corresponds to the edge set representing the social relationships between users. Unlike the traditional social network structure, each edge in *E* is directed and weighted. We define the known local community of the graph as *D* and the partially observable neighbor set of nodes in *D* as *B*. To obtain the entire network *G*, we must visit the neighbor $b(b\in B)$ of the node $v(v\in D)$ constantly. When certain conditions are satisfied, we set b and b 's neighboring nodes to *D* and *B*, respectively. In addition, *D* is usually divided into two subsets:

- The core node set *C*: any node $c\in C$ has no outward connections; that is to say, all the neighbors of c belong to *D*.
- The boundary set *S*: each node $s\in S$ has at least one neighbor in *B*.

The process of local community detection can now be formalized as follows. Given an initial node *v*, a node is added to *D* at each step to discover the target community. Figure 2 shows the sets of nodes *D*, *B*, *C*, and *S* in the social network.

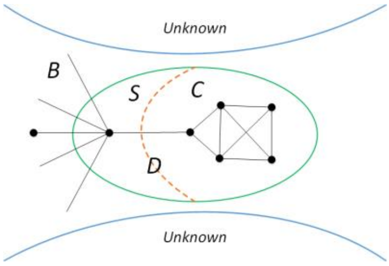

Fig. 2. Definition of the local community

Based on observations of the real social network, we find that social networks have two important features: information sharing and interest diversity. Information sharing means that users tend to share the same content topics in the same community. Interest diversity shows that users often focus on different topics and have social contacts with members in different communities. Traditional methods that always treat *C* as a target community are susceptible to outliers because of the above features. We propose a novel local community detection method setting *D* as the target community using both social relations and topic features.

The following parts of this section present specific interpretations of the proposed algorithm.

3.1 Social Similarity

Social similarity refers to the similarity between a user and a community based on social relations, such as commenting, quoting (@), and retweeting. The higher the social similarity, the higher the probability that the user belongs to the target community. We propose a method for measuring social similarity between user $v(v \in B)$ and community *D*:

$$SS_D(v) = SW_D(v) * SP_D(v)$$

$$SW_D(v) = \frac{1}{2|D|}\sum_{u \in D} \frac{T_{uv}}{K_u} + \frac{T_{vu}}{K_v}$$

$$SP_D(v) = \frac{|\Gamma(v) \cap D|}{|D|} \tag{2}$$

In Eq. (2), $SS_D(v)$ consists of two parts: the social weight $SW_D(v)$ and the social proportion $SP_D(v)$. $SW_D(v)$ represents the sum of social weights between node *v* and all nodes in community *D*. T_{uv} is the number of social interactions from user *u* to user *v*, and K_u is the total number of social interactions of user *u*. T_{vu} and K_v are similar to T_{uv} and K_u, respectively. $SW_D(v)$ takes into account the direction of social connections between users and normalizes the social weights. $SP_D(v)$ is the proportion of the node *v*'s neighbors in community *D*, where $\Gamma(v)$ is the out neighbor set of *v*. The definition of $SS_D(v)$ is similar to term frequency-inverse document frequency (*TF-IDF*) [18] in the field of information retrieval and text mining. Therefore, social similarity has a positive correlation with social weight and social proportion. A user with greater social similarity is more likely to be a member in *D*.

The follower-ship/followee-ship are widely used as social relations in traditional methods, but they have some disadvantages. (1) Excavating the local community from popular users who are followed by many users will likely cause substantial noise. (2) The relationship of following is relatively static and easy to operate in general. To find a stable local community structure, we adopt relative dynamic social relations in the social networks (e.g., quoting, retweeting and commenting) to calculate the social similarity.

3.2 Topic Similarity

"Community Homophily" is when users have the same or similar natures in the same community, such as the same topics in generated content [19]. Inspired by "Community Homophily," we can judge whether an outside user belongs to a particular community by calculating the topic similarity between the user and the community. In general, each piece of generated content is too short to extract topics from; therefore, we aggregate all content from the same user into a document representing that user. The community document is generated by aggregating all users' documents in the community. Then, we calculate the similarity between users and communities using topic distributions, which are extracted from user and community documents by the *LDA* algorithm. In fact, the more similar the topics, the greater the similarity. Typically, Kullback-Leibler divergence is used to calculate the similarity of two distributions. However, we adopt Jensen-Shannon divergence to calculate the topic similarity because the symmetry and triangle inequality cannot be satisfied by Kullback-Leibler divergence. We modify the topic similarity, which is defined in [20], to calculate the topic similarity between user $v(v \in B)$ and community *D* as follows*:*

$$TS_D(v) = 1 - \sqrt{D_{JS}(v, D)} \tag{3}$$

$D_{JS}(v, D)$ is the Jensen-Shannon divergence between two probability distributions and is specifically defined as:

$$D_{JS}(v, D) = \frac{1}{2}\left(D_{KL}(\theta_v \| M) + D_{KL}(\theta_D \| M)\right) \tag{4}$$

θ_v refers to the topic probability distribution of user *v*, and θ_D is community *D*'s topic probability distribution. *M* is the average of two probability distributions:

$$M = \frac{1}{2}(\theta_v + \theta_D) \tag{5}$$

D_{KL} is used to calculate the Kullback-Leibler divergence of probability distributions Q and P:

$$D_{KL}(P \| Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{6}$$

To reduce the computational complexity caused by the dynamic increase of community *D*, we adopt an alternative definition of $TS_D(v)$, as follows:

$$TS_D(v) = \frac{1}{|D|} \sum_{u \in D} TS_u(v) \tag{7}$$

$TS_D(v)$ measures the average topic similarity between user *v* and all users in community *D*. In Eq. (7), *u* is a node in *D*, and $TS_u(v)$ is the topic similarity of user *u* and user *v*.

3.3 Algorithm to Detect Local Communities

In the first two parts of this section, the social similarity and topic similarity between users and communities are defined. Both can play important roles in the aggregation of community, and thus, the fusion similarity is proposed as follows:

$$FS_D(v) = \gamma SS_D(v) + (1-\gamma) TS_D(v) \tag{8}$$

In Eq. (8), γ is the fusion coefficient that balances the proportions of social similarity and topic similarity. In the local community detection algorithm, the node *v* with the highest $FS_D(v)$ will be selected to be aggregated into *D* after each update. Usually, researchers are more concerned with the community network composed of the top *N* nodes sorted by fusion similarity to obtain fewer outliers and high precision.

By adding one node to set *D* at each step, our algorithm can explore a target community from an initial node. In each step, the algorithm calculates the social similarity and topic similarity simultaneously. Moreover, the algorithm can also be applied to explore more complete community structures using multiple initial nodes identified from the same community.

4 Experiment and Evaluation

To evaluate the performance of the proposed local community detection method, we employ the real Micro-Blog datasets. As a platform for information acquisition and sharing, Weibo (a micro-blog platform similar to Twitter) has become one of the most popular social networks in China and has significant community characteristics. Users in the same community are closely connected and share common interests, whereas connections between different communities are relatively loose, and their interests are often very different. The above characteristics are consistent with Newman and Girvan's formal definition of community [1].

4.1 Dataset Preparation

The data is crawled using the public API[1] of Weibo, and the preprocessing steps are as follows:

Remove the users with fewer than 10 micro-blogs: This step was performed because these users have a small number of micro-blogs and social relationships to be extracted.

Extract social relationships. For all the micro-blogs of each user, we found social relations, such as "@" and "//@:", using regular expressions and extracted the root users from forwarding or comment micro-blogs. A social relation matrix was formed after screening and statistics.

Preprocess the micro-blog text content: We put all of a user's aggregated micro-blogs into a document (i.e., forming a one-to-one relationship between the user and the document). We first removed the "@" and "//@:" behaviors from the texts and then used the Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS)/Natural Language Processing Information Retrieval (NLPIR) [21] to cut documents into words and remove any nonsense words, symbols, and expressions. Finally, all of the documents were aggregated into one, and the mapping relationships between users and documents were saved.

Table 1 tabulates the statistics of the final experimental dataset. The users in the experimental dataset covered technology, politics, the economy and other fields, thus providing our method with broader applications.

According to [15-16], the *LDA* parameters were set as follows: topic number $T = 50$, $\alpha = 50/T$, $\beta = 0.1$, and iteration number=1000. These are empirical parameters used to produce better results after many tests.

[1] http://open.weibo.com/wiki/

Table 1. Statistics of the dataset used for evaluation

Category	Amount
Number of total users	1206
Number of micro-blogs	1248071
Number of non-repeating items in micro-blog texts	293837
Number of social relations	157685
Average social interactions of users	130

4.2 Evaluation Criterion

Weibo provides a function that shows similar users on the homepage. For a user, we name his/her similar users as first-order-similarity users, and his/her first-order-similarity users' similar users are called second-order-similarity users. We crawled the seed user's first-order-similarity and second-order-similarity users as the candidates. We invited three experts to screen the candidate users manually. By considering the candidates' basic information, social rules, interest topics and other attributes, *M* users were selected as members of the local community of the seed user.

The evaluation metrics we used—precision (*P*), recall (*R*) and *F1-Score*—are quite simple and are quite frequently used in many areas, such as information retrieval and machine learning. Other papers focusing on the community detection problem have also adopted these metrics [6, 8, 10]. The size of the detected local community is *N*, and *A* is the correct number of users in the community. These metrics were calculated using Eq. (9):

$$P=\frac{A}{N},R=\frac{A}{M},F1=\frac{2P*R}{P+R} \tag{9}$$

4.3 Experimental Results

In our experiments, we compared the results of different local community detection methods. Descriptions of the labels we used to denote each of these algorithms are presented below:

- **Clauset.** This is a basic algorithm defining the local modularity R [4].
- **LWP.** This is an improved two-phase algorithm that defines a new local modularity M [5].
- **Chen.** Chen proposed another method for local community detection that defines the metric L to evaluate the local community structure [6].
- **LS-M.** This method is a version that uses link similarity with local modularity M [8]
- **LS-R.** This method is another version that uses link similarity with local modularity R [8]
- **S-LCD.** This is a method that we proposed for local community detection based only on social relations.

- **T-LCD.** Similar to the above, this performs local community detection based only on topic features.
- **F-LCD.** Similar to the above, this performs local community detection based on both social relations and topic features.

A user was randomly selected as the seed in the data set; that user's ID was #130. The user is a researcher in the field of artificial intelligence through the observation of profiles.

Comparing the precision with traditional local detection methods

We first compared the precision with those of traditional local detection methods. Figure 3 shows the results of different algorithms in terms of their precision. Our algorithm, which is based on both social relations and topic features, performs the best. In this experiment, the community size N was 100, and the fusion coefficient γ was set to 0.8. In the following experiments, we compare the influence of different effects of N and γ .

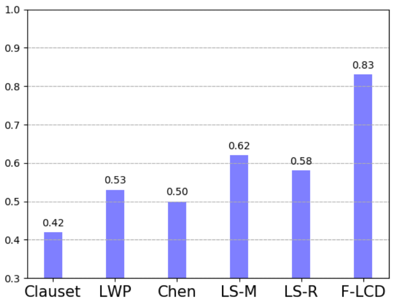

Fig. 3. Precision of different methods

Comparing the precision, recall, and F1-Score of our proposed methods

The community size N was set to 100 and γ was set to 0.8. Figure 4 shows the specific experimental results.

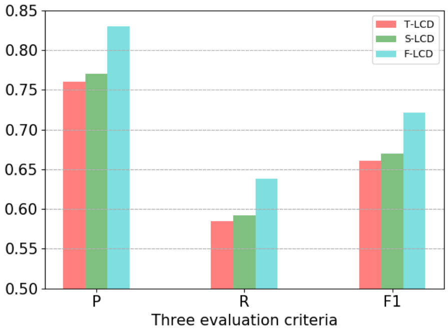

Fig. 4. Precision, recall, and F1-Score of different methods

As shown in Figure 4, the local community detection method based on the fusion of social relations and topic features performs better than the other two methods in terms of their precision, recall and F1-Score. Local community detection based on social relations alone focuses on closely connected users within the community, but according to interactive topics, many users may not belong to the local community. In contrast, the approach that relies on topic features alone selects the most similar users based on content topics as local community members, and the internal links within the community are usually very sparse and do not conform to the formal definition of a community.

Table 2 shows the top 20 words from the highest probability topic in the distribution θ of the target community. Translating the Chinese words in the micro-blogs into English reveals that the users in the target community are similar to the seed user in terms of their topic words.

Table 2. Words with the highest topic probability in the distribution

Index	Word	Index	Word
1	Learning	11	System
2	Data	12	Artificial Intelligence
3	Machine	13	Microsoft
4	Thesis	14	Field
5	Research	15	Work
6	Deep	16	Model
7	Technology	17	Algorithm
8	Problem	18	Recommendation
9	Paper	19	Compute
10	Professor	20	Schoolmate

Effects of community size N and fusion coefficients γ on precision, recall, and F1-Score

Community size N and fusion coefficients γ are important parameters, and different values of them will affect the final results. In this part, the experimental results obtained using different values of parameters are presented.

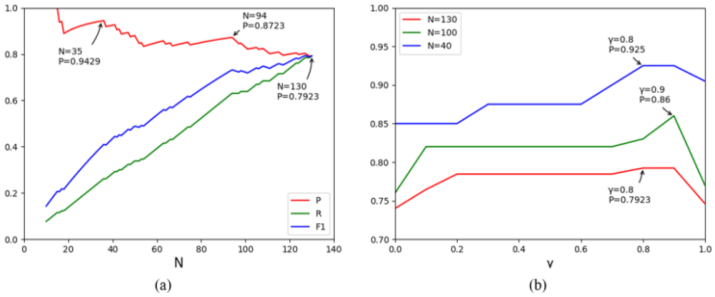

Fig. 5. (a) Comparison of precision, recall, and F1-Score for different community sizes; (b) Comparison of the precision for different fusion coefficients γ

Figure 5 (a) compares the precision, recall, and F1-Score for different community sizes. The *F-LCD* algorithm is adopted in this part, and the fusion coefficient γ is set to 0.8. As *N* increases, the precision curve shows a decreasing trend. However, when *N* equals 35 or 94, the precision reaches a local optimal value. In fact, when *N* is equal to approximately 100, the *F-LCD* algorithm can achieve a precision of more than 83% after many tests, which is an acceptable result in practice. In addition, the recall and F1-score increase as *N* increases.

Figure 5 (b) shows the effects of different fusion coefficients γ on the precision. We use different values of *N* to implement the experiment to obtain the general rule. When γ is set to 0 or 1, the *F-LCD* method is equivalent to the *T-LCD* or *S-LCD* method. The precision is the highest when γ is between 0.8 and 0.9 for different values of *N*.

The community size *N* and fusion coefficient γ are important parameters in the local community detection algorithm. The last experiment compared the results obtained by using different *N* and γ in the *F-LCD* algorithm. People always want to get the maximum benefit with some prior knowledge. Therefore, the experimental results have significance for the practical application of this method.

5 Conclusions

Community detection is an important research focus in social network analysis. Currently, local community structures can only be detected from a given seed node because of the incomplete network structure and high computational complexity. Traditional local methods are mainly based on undirected and unweighted networks, and thus, such methods have limitations when applied to popular social networks. This paper defines a new social similarity metric to measure the link similarity between a user and a local community using directed and weighted relations. Then, the classical Jensen–Shannon divergence is used to calculate the topic similarity, in which the topics are extracted from the user's text content by topic models. Finally, a novel algorithm based on the fusion of social relations and topic features is proposed for local community detection. Extensive experiments on a real social network dataset demonstrated the efficacy of the proposed algorithm. Because the networks are studied offline, one possible direction for future work is to discover the dynamic online communities and analyze their evolution processes.

References

1. Newman M E, Girvan M. Finding and evaluating community structure in networks.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2004, 69(2 Pt 2):026113-026113.
2. Mislove A, Marcon M, Gummadi K P, et al. Measurement and analysis of online social networks[J]. 2015.
3. Newman M E J. Detecting community structure in networks[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2004, 38(2): 321-330.

4. Clauset A. Finding local community structure in networks[J]. Physical review E, 2005, 72(2): 026132.
5. Luo F, Wang J Z, Promislow E. Exploring local community structures in large networks[J]. Web Intelligence and Agent Systems: An International Journal, 2008, 6(4): 387-400.
6. Chen J, Zaïane O, Goebel R. Local community identification in social networks[C]//Social Network Analysis and Mining, 2009. ASONAM'09. International Conference on Advances in. IEEE, 2009: 237-242.
7. Bagrow J P, Bollt E M. Local method for detecting communities[J]. Physical Review E, 2005, 72(4): 046108.
8. Wu Y J, Huang H, Hao Z F, et al. Local community detection using link similarity[J]. Journal of computer science and technology, 2012, 27(6): 1261-1268.
9. Zhang T, Wu B. A method for local community detection by finding core nodes[C]//Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society, 2012: 1171-1176.
10. Chen Q, Wu T T, Fang M. Detecting local community structures in complex networks based on local degree central nodes[J]. Physica A: Statistical Mechanics and its Applications, 2013, 392(3): 529-537.
11. Lancichinetti A, Fortunato S, Kertész J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2008, 11(3):19-44.
12. Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2003, 101(9):2658-2663.
13. Papadopoulos S, Skusa A, Vakali A, et al. Bridge Bounding: A Local Approach for Efficient Community Discovery in Complex Networks[J]. Physics, 2009, 1(1-12):174.
14. Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks.[J]. Nature, 2010, 466(7307):761-4.
15. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3:993-1022.
16. Griffiths T L, Steyvers M. Finding scientific topics.[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 Suppl 1(1):5228-35.
17. Minka T, Lafferty J. Expectation-Propagation for the Generative Aspect Model[J]. Journal of Computational & Applied Mathematics, 2002, 235(11):3257-3269.
18. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval ☆[J]. Information Processing & Management An International Journal, 1988, 24(5):513-523.
19. Bisgin H, Agarwal N, Xu X. Investigating Homophily in Online Social Networks[C]// Ieee/wic/acm International Conference on Web Intelligence, Wi 2010, Toronto, Canada, August 31 - September 3, 2010, Main Conference Proceedings. 2010:533-536.
20. Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic-sensitive influential twitterers[C]// International Conference on Web Search and Web Data Mining, WSDM 2010, New York, Ny, Usa, February. 2010:261-270.
21. Zhang H P, Yu H K, Xiong D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]// Sighan Workshop on Chinese Language Processing. Association for Computational Linguistics, 2003:págs. 758-759.