doi:10.3772/j.issn.1006-6748.2018.02.010

Research on community detection algorithm based on site topic similarity and topology[①]

Hu Yi(胡　艺)[*], Li Zhengmin[**], Chi Lejun[②][*], Lin Jinxiu[*]
([*]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P. R. China)
([**]National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, P. R. China)

Abstract

Considering the deviation in content of community detection resulting from the low accuracy of resource relevance, an algorithm based on the topology of sites and the similarity between their topics is proposed. With topic content factors fully considered, this algorithm can search for topically similar site clusters on the premise of inter-site topology. The experimental results show that the algorithm can generate a more accurate result of detection in the real network.

Key words: site relationship network, community detection, topic similarity, clustering

0 Introduction

System science holds that structure determines functions. A network system consists of a considerable number of nodes and inter-node relationships, and such system structure is known as a complex network[1]. A site relationship network, studied in this paper, is a complex network, which can be represented by a graph, where the sites are the nodes and the inter-site linkages are the edges.

A complex network includes small-world effects[2], scale-free characters[3], and node type diversity[4]. The community detection in a complex network is a process of structure-based metric partition, which can help people understand the hierarchical relationships in a network, and reveal the latent attributes among these relationships[5].

Community detection for site relationship network aims to reveal the resource sharing attributes of sites, so that it could be available for collaborative recommendations to achieve the goal of knowledge sharing.

The main idea of common community detection algorithms, such as modularity and min-cut[6], is to identify the relationship of strength among network edges. But modularity algorithm has been proved to have limitations in resolution ratio at present: when modularity is applied to optimization, the modules whose size is smaller than a certain value cannot be parsed[7]. Moreover, since the similarity of node attributes cannot be effectively guaranteed without the node content attributes, the modularity algorithm does not meet the requirements of community detection for site relationship network well.

Community detection based on topic partition, however, cannot reveal inter-node relevance, and any two nodes are not necessarily closely relevant to each other even if they share the same topic.

According to the implementation requirements of a relevance knowledge recommendation system with sites as objects, a comprehensive analysis needs to be carried out on site topic relevance and inter-site link attributes. Therefore, in the community partition for site relationship network, both site topics and inter-site links are factors to be considered.

In view of the above, this paper presents a community detection algorithm based on site topic similarity measurement and network topology. First, n topic tags from the site content are extracted. Then, the modularity gain and the topic similarity gain are used to evaluate whether the two nodes should be grouped into the same community.

The rest of this paper is organized as follows. In Section 1 some related efforts on community detection have been discussed. Section 2 presents the algorithm proposed in this paper. Experimental results are given in Section 3. Finally, the paper is concluded in Section 4.

[①] Supported by the National Science and Technology Support Program of China (No. 2012BAH45B01), the National Natural Science Foundation of China (No. 61100189, 61370215, 61370211, 61402137) and the National “242” Project of China (No. 2016A104).

[②] To whom correspondence should be addressed. E-mail: qdclj@163.com

Received on June 1, 2017

1 Related work

Highly time-efficient community detection algorithms include Louvain algorithm and Label Propagation, etc. As an improvement on modularity, Louvain algorithm is applied to large-scale networks, and is well received due to its high efficiency as well as good community quality[8]. In recent years, lots of researchers have improved and optimized the convergence criteria of Louvain algorithm variously in order to gain higher time efficiency and modularity, so that it could be widely used in large-scale datasets[9-11].

However, the currently improved Louvain algorithm still takes modularity as a treatment factor, while edge weight has a comparatively single dimension, and there is a big difference in topic tendency between the partition results.

In terms of topic combination, the present research mainly focuses on microblog. Lu, et al.[12] selected topics from short texts by latent topic modeling, and then implemented large-scale microblog clustering by two-layer hybrid clustering. Yan, et al.[13] calculated the shortest path length and directional relationship in microblog, and took their weighted average as a weight factor of community detection for the implementation of community detection. This method achieved a good result in community topic purity. Sun, et al.[14] defined the modularity function of similarity based on microblog users' common attention and following behavior, and they performed community detection by greedy algorithm. Wang, et al.[15] extracted user topics by LDA model, then filtered cluster centers in accordance with the path distance, and finally mined microblog communities based on topic similarity. This method can be used to gather latent communities and topics, but the number of communities is limited by the clustering variables set manually. Li, et al.[16] proposed a clustering approach based on hierarchical expansion, which built the scientific collaboration network to calculate the strength of the connected between each two nodes, and used modularity Q to determine the quality of the cluster, which provided a new idea for the community detection and exploitation of scientific collaboration network.

Unlike microblog where there are just short texts, a site page contains all sorts of documents, and inter-site relationship mainly consists of link orientations. Therefore, considering the characteristics of site relational network, analyzing the advantages of the existing microblog clustering schemes, and by reference to the convergence criteria of efficient Louvain algorithm, this paper proposes an algorithm based on topic similarity and topology——TSTA (algorithm based on topic similarity and topology). This algorithm first extracts site labels, then defines the evaluation function of comprehensive label similarity and modularity, and finally searches for an approximately optimal community partition resulting in combination of greedy algorithm.

2 TSTA implementation

In this section, an algorithm (called TSTA) for detection of communities in graph is proposed. The data model is first proposed for the algorithm in subsection 2.1 below, a computational scheme is designed for topic similarity in subsection 2.2, and the detailed steps of algorithm are described in subsection 2.3.

2.1 Site relationship network data modeling

Site relationship network can be represented as graph G with a set of nodes N, a set of edges E and a set of tags Z. The nodes represent the sites in a real-world network whereas edges can be considered as the link relations between the sites, the tags are a collection of topics keywords. Graph $G = (V, E, Z)$ is as shown in Fig. 1, where:

1) V is the set of node members in the graph, $V = \{v_1, v_2, v_3, \cdots, v_n\}$;

2) E is the set of edges, $E = \{e_1, e_2, e_3, \cdots, e_n\}$, with $m = | E |$, here, the density of the link as the weight, while the link buffer value and no follow attribute are not considered temporarily;

3) Z is the set of topic (or content) tags for members, which is extracted from the site document T, $Z = \{z_1, z_2, \cdots, z_m\}$, $T = \{t_1, t_2, \cdots, t_m\}$.

4) Similar (i, j) represents the similarity between v_i and v_j two node tags.

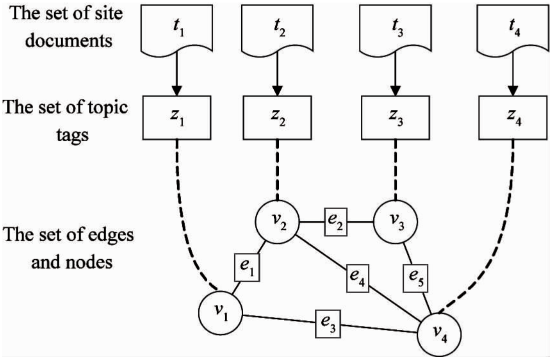

Fig. 1 The data models for TSTA

The output of the algorithm is the set of communities $C = \{C_1, C_2, \cdots, C_k\}$, where:

1) $w(\sum_{inC_j})$ is the sum of all the weights in C_j;

2) $w(\sum_{toC_j})$ is the sum of the weights of the nodes in all incident C_j;

3) w_{i,C_j} is the sum of the weights from node v_i to C_j.

2.2 Site topic introduction

Site content is a set of multiple html documents. Content vocabulary set T is first chosen from site documents in this paper, and then documents are processed by model TF-IDF (term frequency-inverse document frequency), forming topic document z. TF-IDF is a derivative of statistical thoughts, which is also applicable to the text sets composed of various site documents since it can be used to judge whether a word has a strong labeling ability by calculating the exposure frequency of this word in the current document and other documents. In this paper, words are sorted by the size of TF-IDF value, and the first m words are extracted as labels of site A, $A_z = \{A_{z1}, A_{z2}, \cdots, A_{zm}\}$.

After a set of site labels acquired, cosine similarity algorithm, as in Eq. (1), is introduced into this paper to evaluate the similarity between site label A and B, of which the value range between 0 and 1, and a higher value means that the two sets of labels are more similar to each other.

$$similar(A,\ B) = \frac{\sum_{k=0}^{m-1} A_{zk} \times B_{zk}}{\sqrt{\sum_{k=0}^{m-1} {A_{zk}}^2} \times \sqrt{\sum_{k=0}^{m-1} {B_{zk}}^2}} \tag{1}$$

The S of a community is a scalar value between −1 and 1 that measures the mean value of the similarity between every two nodes in the community, as in Eq. (2), where n represents the numbers of nodes in community C_m. The larger the value is, the higher the topic quality is.

$$S = \frac{1}{k} \sum_{m=1}^{m=k} \frac{\sum_{i,\,j \in C_m} similar(i,\ j)}{C_n^2} \tag{2}$$

2.3 Detailed description of TSTA

The community detection algorithm proposed in this paper is divided in two phases: the phase of node mobility and the phase of node combination.

At the phase of node mobility, first, the gain generated by placing node i in the community of each neighbor j, is calculated, which is the weighted average of the degree of variance of modularity gain and topic similarity, as shown in Eqs (3) and (4), then the node i selects the neighbor with the highest gain to merge. This phase is circulated until all gains get less than 0.

At the phase of node merge, the nodes falling under the same community at the above phase are primarily reshaped as independent nodes in the merging process of edge weight and topic label, etc., namely new graph $G'(V',\ E',\ Z')$, with $V' = \{C_n, C_m, \cdots, C_k\}$, $w'_{i,j} =$ (where $v_i \in C_i, v_j \in C_j$), the new node topic label $Z'_n = \sum_{inC_m} z_i$.

These two phases are iterated until there are no longer grow of modularity Q, as in Eq. (5).

$$\Delta P = \lambda \times \Delta Q + (1 - \lambda) \times \Delta Q \times similar(i,\ j) \tag{3}$$

$$\Delta Q = \left[\frac{w(\sum_{inC_j}) + w_{i,C_j}}{2m} - \left(\frac{w(\sum_{coC_j}) + w_i}{2m}\right)^2\right] - \left[\frac{w(\sum_{inC_j})}{2m} - \left(\frac{w(\sum_{coC_j})}{2m}\right)^2 - \left(\frac{w_I}{2m}\right)^2\right] \tag{4}$$

$$Q = \sum_{C_i \subset C}\left[\frac{\sum inc_i}{2m} - \left(\frac{\sum coc_i}{2m}\right)^2\right] \tag{5}$$

According to the design of the TSTA implementation above, Algorithm 1 shows details of gain calculation part and Algorithm 2 represents the pseudo code of TSTA steps.

Algorithm 1 maximum _ gain _ calculation

Input: node

Output: best _ gain, best _ neighbor

```
Begin
for random $v \in neighbor(node)$ do
    gain ← $\triangle P$ from move (node, community(v))
    if gain > best _ gian then
        best _ gain ← gain
        best _ neighbor ← v
Return best _ gain and best _ neighbor
END
```

Algorithm 2 TSTA

Input: A graph $G = (V, E, Z)$

Output: A set of communities $C = \{C_1, C_2 \cdots\cdots C_k\}$

```
Begin
  $Q = 0$
  network[0] ← initialize(G, network[0])
Do {
    Best _ Q = Q
    Do{
        improvement ← false
        for node in network[0]:
            $P_{gain}$, target _ neighbor←
                maximum _ gain _ calculation(node)
            if $P_{gain} > 0$ then
                merge the node with target _ neighbor
                improvement ← true
        } while(improvement)
        rebuild(network[0])          //merge node
        q ← calculate Q (network[0])
    } while q > = best _ Q
    Return network[0]
END
```

3 Experimental results and analysis

The algorithm proposed in this paper is validated with real network data in this section. The experimental environment of this paper is described in subsection 3.1. Subsection 3.2 presents the evaluation indicators involved in the experiment. Subsection 3.3 compares the experimental results of this algorithm with the traditional modularity algorithm.

3.1 Experimental environment

Experimental datasets: more than 40 miscellaneous sites are adopted in this paper as seed sites (including game, training, news and other relevant topic sites), and then with link relation as a relevance factor, over 8,000 sites are selected as nodes for relevance analysis.

Working environment: the experiment is con-ducted by a desktop computer (1.6GHz Intel Core i5 processor, 4GB memory, 1600 MHz DDR3).

3.2 Experimental methods and evaluation indexes

Considering that the algorithm presented in this paper is based on topic similarity and modularity, the evaluation values containing these two factors are selected.

First, the difference in community topic expression between Louvain and TSTA is analyzed based on community topics in partition results. The followings are evaluations made from the perspective of topic distribution ratio, topic looseness and highest correlation proportion.

1) Topic distribution ratio: Topic distribution ratio reveals the distribution of topics in target community, and if there is some topic whose proportion is greater than the threshold (50% here), it suggests that this topic is significantly representative.

2) Topic looseness: Topic looseness refers to the number of communities that contains topic A. If there are n communities in the partition result that contain topic A, the looseness of topic A in this partition result is n. The higher the looseness is, the more dispersed the topic distribution is.

3) Best relevance ratio: The number of sites containing topic A in the community where topic A appears the most frequently is denoted by n_A, and the number of the sites containing topic A is n, and the best relevance ratio is n_A/n. The larger the best relevance ratio is, the more concentrated the topics are, namely the more likely there is the topic clustering required.

After the above three factors are analyzed, this paper combines modularity with topic similarity, completing the final evaluation based on an objective value. Here, inspired by F-source, and with reference to the evaluation approaches in Refs[17,18], a calculation formula is established for value of evaluation by combining S value with Q value, as shown in

$$SQ\text{-}F\alpha = \frac{(1+\alpha^2)\cdot(S\times Q)}{\alpha^2\cdot(S+Q)} \qquad (6)$$

where α is the weight parameter to adjust S and Q, $\alpha \in [0, \infty)$ when $\alpha = 1$, S and Q have the same weight. The larger the value of $SQ\text{-}F\alpha$ is, the better the result of weighing by both factors is.

3.3 Experimental process and results

At the first step, this paper preprocesses the datasets, and takes the home information on each site as its content information. First, Chinese sentences in the html documents are fragmented and cleaned, with null words as well as classifier, adverb and temporal phrases deleted, and the result of each site is output as preprocessing data. Then, the keywords with strong labeling ability are output by model TF-IDF model as site labels and brought into the algorithm processing procedure. The above preparation step is completed.

At the second step, community testing is performed by TSTA, and Louvain is adopted as the comparison object.

1) Community mining results

The TSTA and Louvain algorithms are used to process the datasets. In the station relation network of 1000, 2000, 3000 and 4000 levels, the result shown in Fig. 2 is obtained.

As can be seen from the experimental data, with the number of sites increased progressively, the number of communities discovered by Louvain tends to be stable, and easy to include large-scale communities. The improved algorithm discovers much more sites than the former algorithm, and the average size of communities is relatively stable.

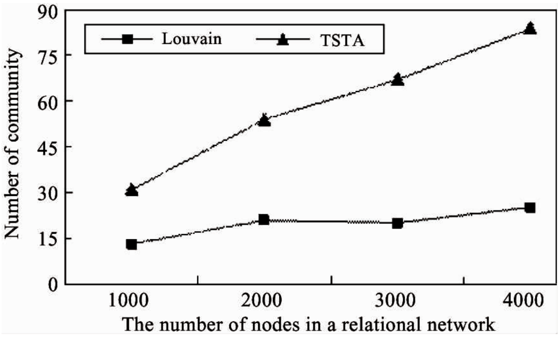

Fig. 2 The number of communities of different network sizes.

In order to further verify the algorithm presented, this paper takes 1,000 site relationship networks as a dataset (Fig. 3), and makes a more detailed analysis on the experimental results.

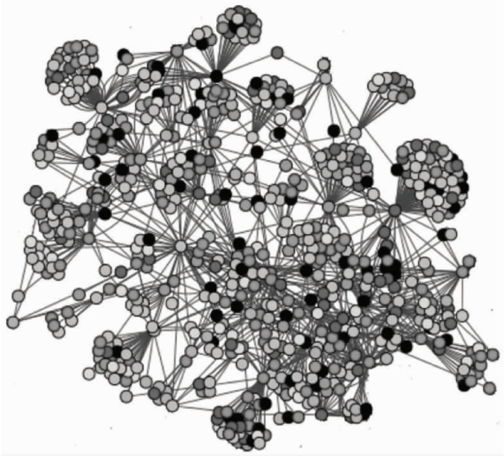

Fig. 3 Experimental data set

Fig. 4 shows the community partition result by Louvain (with independently scattered points excluded). Extracting and analyzing the topic labels of any two communities which are close to the average in the result, get a scale shown in Fig. 5. It can be seen that there are multiple sites with different topic labels in the same community, and there is no topic whose ratio is greater than the threshold, which reflects the looseness of the topic characteristics of the individual community.

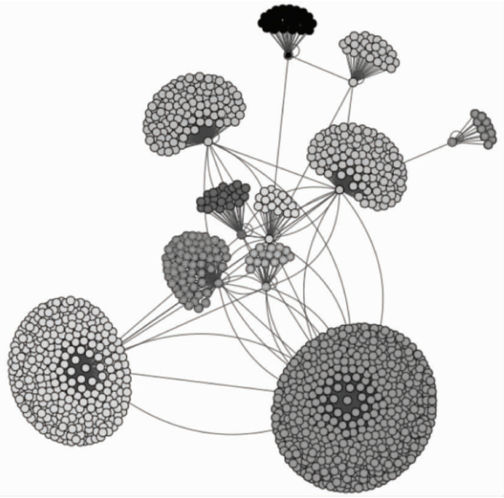

Fig. 4 The detected communities in the network by Louvain.

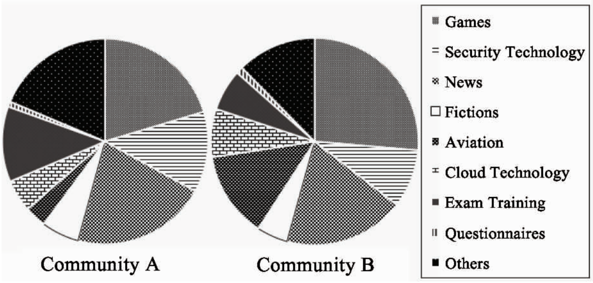

Fig. 5 Scale map of topic distribution under Louvain.

Then, a comparative analysis is performed on the result of TSTA, and Fig. 6 shows the result of community partition by TSTA.

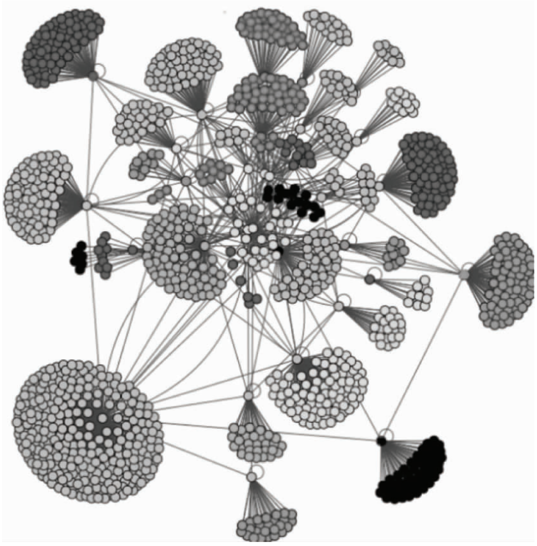

Fig. 6 The detected communities in the network by TSTA

1,000 datasets are partitioned into 32 communities by TSTA. Fig. 7 is the topic scale distribution in two of the communities. As can be seen, there are topics whose ratio is greater than 50% in these two communities, namely there has been a topic tendency in them.

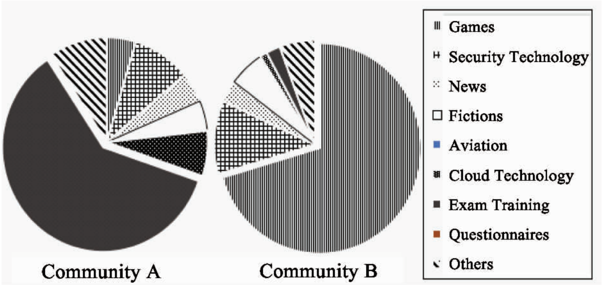

Fig. 7 Scale map of topic distribution under TSTA

in them. However, since the algorithm couldn't break the inter-site connections in principle, absolute topic centralization couldn't be realized.

Fig. 8 is the difference in topic looseness between Louvain and TSTA, and Fig. 9 is the best relevance ratio. As can be seen, the mean topic looseness by TSTA is lower than that by Louvain, while the best relevance ratio is larger than the latter. The topic convergence expected could be seen in the community detection result, which is more aligned with the requirements of community detection in site relationship networks.

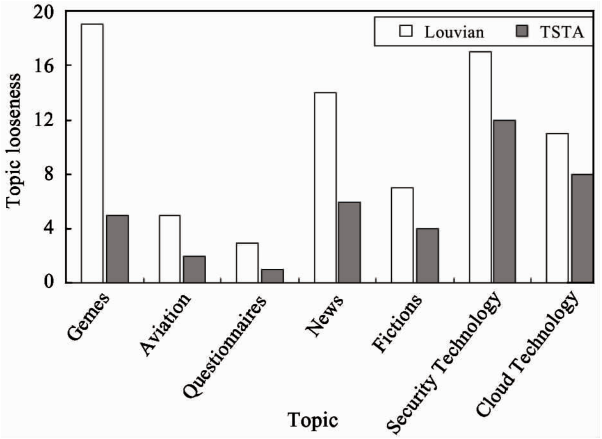

Fig. 8 Topic looseness for Louvain and TSTA

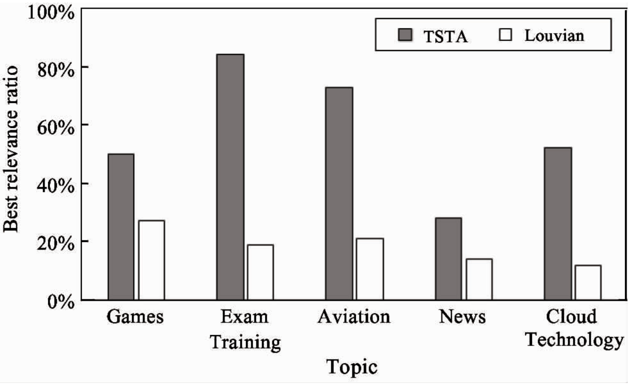

Fig. 9 Best relevance ratio for Louvain and TSTA

2) SQ-Fα

Then the SQ-Fα metric is used to evaluate the results of TSTA and Louvain, and parameter α is taken as a varying factor set to 0.5, 0.75, 1, 1.25 and 1.5 respectively, with a relation curve shown in Fig. 10. It can be seen that in the real network, the SQ-Fα-based results of the TSTA are better than those of the modularity algorithm. In other words, TSTA has better performance than the Louvain, as well as a more distinct advantage in integrated modularity and topic quality.

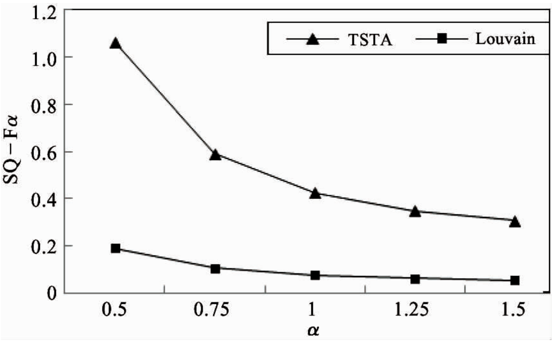

Fig. 10 SQ-Fα for Louvain and TSTA on different α

At the same time, the experiment finds that the better community detection result obtained by TSTA algorithm is based on longer calculation time. In theory, the core computing path of TSTA algorithm is basically the same as Louvain algorithm, so that its time-complexity is the same as Louvain algorithm. However, since TSTA algorithm introducing cross-service calculation, which adds network and cosine similarity model calculation besides basic time, so the real working time is 5-7 times of Louvain algorithm with the same number of nodes.

4 Conclusion

This paper studies the community detection technologies for site relationship network, analyzes the shortcomings of common community detection algorithms, and presents a community detection algorithm based on topic similarity and topology. The proposed method first performs site topic similarity matched by TF-IDF topic modeling and cosine similarity algorithm, and then strives to gain optimal group benefits in community detection using the concept of modularity and similarity gain. The results of experiments and comparisons with the traditional modularity algorithm indicate that our algorithm performs a greater advantage both in modularity and topic quality , thus it can be used to mine related site resources.

References

[1] Dame N. Statistical mechanics of complex networks[J]. *Review of Modern Physics*, 2002, 74(1):xii

[2] Watts D J, Strogatz S H. Collective dynamics of small world networks[J]. *Nature*, 1998, 393(4): 440-442

[3] Barabási A, Albert R. Emergence of scaling in random networks[J]. *Science*, 1999, 286(5439): 509-512

[4] Newman M. Networks: An Introduction[M]. New York: Oxford University Press, 2010. 2-11

[5] Gan W Y, He N, Li D Y, et al. Community discovery method in networks based on topological potentia[J]. *Journal of Software*, 2009, 20(8): 2241-2254

[6] Wang L, Dai G Z. Community finding in complex networks——theory and applications[J]. *Science & Technology Review*, 2005, 23(0508): 62-66

[7] Fortunato S, Barthélemy M. Resolution limit in community detection[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, 104(1): 36-41

[8] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. *Journal of Statistical Mechanics Theory & Experiment*, 2008, 2008(10):155-168

[9] Traag V A. Faster unfolding of communities: speeding up the Louvain algorithm[J]. *Physical Review E*, 2015, 92(3): 032801

[10] Gach O, Hao J K. Improving the Louvain Algorithm for Community Detection with Modularity Maximization[M]. In: Proceedings of the Legrand P, Corsini M M, Hao J K, et al. (eds) Artificial Evolution, Springer, 2013. 145-156

[11] Wu W J, Li M N, Li G H. Parallel processing of the Louvian algorithm[J]. *Computer & Digital Engineering*, 2016, 44(8): 1402-1406

[12] Lu R, Xiang L, Liu M R, et al. Discovering news topics from micorblogs based on hidden topics analysis and text clustering[J]. *Pattern Recognition and Artificial Intelligence*, 2012, 25(3): 382-387

[13] Yan G H, Shu X, Ma Z C, et al. Community discovery for microblog based on topic and link analysis[J]. *Application Research of Computers*, 2013, 30(7): 1953-1957

[14] Sun Y F, Li S. Similarity-based community detection in social network of microblog[J]. *Journal of Computer Research and Development*. 2014, 51(12): 2797-2807

[15] Wand W P, Fan T. Community discovery method based on users' interset similarity and social network structure [J]. *Computer Systems & Applications*. 2013, 22(6): 108-113

[16] Li X H, Zheng Y N. Clustering approach based on hierarchical expansion for community detection of scientific collaboration network [J]. *High Technology Letters*, 2016, 22(4):419-425

[17] Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering[C]. In: Proceedings of the Workshop on Artificial Intelligence for Web Search (AAAI 2000), Austin, USA, 2000. 58-64

[18] Zhao Z, Feng S, Wang Q, et al. Topic oriented community detection through social objects and link analysis in social networks[J]. *Knowledge-Based Systems*, 2012, 26: 164-173

Hu Yi, born in 1993. She received her M. S. degrees in computer technology from Harbin Institute of Technology. Her research interests include the network and information security, open source intelligence.