TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.5, May 2014, pp. 4038 ~ 4049
DOI: http://dx.doi.org/10.11591/telkomnika.v12i5.4170

Community Detection Based on Topic Distance in Social Tagging Networks

Hongtao Liu, Hui Chen*, Mao Lin, Yu Wu
School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065
*Corresponding author, e-mail: chenhui88happy@gmail.com

Abstract

Research on the community detection in social tagging networks has attracted much attention in the last decade. Extracting the hidden topic information from tags provides a new way of thinking for community detection in social tagging networks. In this paper, a topic tagging network by extracting several topics from the tags through using the Latent Dirichlet Allocation (LDA) model is built firstly. Then a topic distance between users is defined, which depends on the bookmarking relationships between users and tags. Further, a modularity clustering approach based on the topic distance is proposed to detect communities in social tagging networks. Empirical studies on real-world networks demonstrate that the proposed method can effectively detect communities in tagging networks.

Keywords*: community detection, tagging networks, topic distance, modularity optimization*

Copyright © 2014 Institute of Advanced Engineering and Science. All rights reserved.

1. Introduction

Community detection provides an important way to further understand and apply social networks, that is, identifying communities or groups in which nodes are densely connected inside while loosely connected outside through some methods. As one successful kind of social networks, social tagging networks such as Del.icio.us, CiteULike and Flickr have developed fastly in the past several years. In social tagging networks, users can easily use tags to organize, share and retrieve online resources, which has different meaning in different environments, for example, Web pages in Del.icio.us, research papers in CiteULike and photos in Flickr. At the same time, those tagging networks have created large amounts of tagging data which have attracted more and more researchers to pay attention to identify potential community structure in social tagging networks.

At present, there are many different community detection approaches, and two kinds among them are applied more [1]. The first is traditional topology-based community detection approach, which maps the real world network into a graph structure with nodes representing users in real world network and edges representing the interaction relation between users. Community detection approach based on graph partitioning or clustering tries to detect subgraphs with high density, such as GN algorithm [2], K-L algorithm [3], the spectral bisection method [4] and so on. But those methods mostly research on the structural properties of community, ignoring other important characteristics, especially the theme characteristics of community, for example, users belong to a community tend to have similar hobbies, social function, occupation, interest or viewpoint on the same topic and so on. Therefore, another topic-based community detection methods has been widespread concern, that is, considering the text information in networks and detecting communities according to the content users published. Hierarchical clustering based on distance or similarity metrics is a common one. The topic-modeling approach [5] is another topic-based community detection approach, such as Latent Dirichlet Allocation (LDA) [6] and its various variations, for example, Author-Topic model [7], Community-User-Topic model (CUT) [8], Topic-User-Topic model (TUCM) [9] and so on. And discovering communities consisting of similar users is an important problem and can find practical applications in sociology, biology, computer science and other areas. There has been some related work about how to define the similarity between users based on which people are grouped into communities. One common approach is to treat communities as group of nodes in

Received October 15, 2013; Revised December 28, 2013; Accepted January 17, 2014

social network that the connection among themselves are more densely than with the rest of network, which makes the community detection a graph clustering problem. In Sachan's work [9], communities are considered as "groups of users (nodes) who are interconnected and communicate on shared topics". This paper follows the viewpoint about communities.

In this paper, we propose a community detection approach based on topic distance which is short for TDSHRINK, combining topic information with modularity clustering approach. To summarize, this work contributes on the following aspects: (1) We define a topic distance based on the bookmark relationship between users on same topics. (2) A new modularity clustering algorithm based on topic distance in social tagging networks is proposed, that is, grouping two users into a community if they are interested in the same topics. (3) The algorithm we proposed can detect overlapping communities, that is, a user is allowed to belong to multiple communities. And the accuracy and efficiency of our algorithm are improved compared with other methods based on modularity.

This paper is organized as follows. We introduce the related work on LDA model and modularity in Section 2. The formal definition is presented in Section 3. In Section 4, we describe the topic distance-based modularity clustering algorithm specifically. The experimental results are presented in Section 5. Finally, we conclude in Section 6.

2. Related work

Our work is related to two research areas: LDA model to extract latent topics and community detection approach based on modularity optimization.

2.1. Topic Model LDA

In real life, we always want to find a brief description or summary to represent or reflect the feature information of a large-scale dataset. For example, extracting several topics to represent the total text dataset, while the topic model is the model which can effectively analyze large amounts of text [10]. The most widely used in the topic model is LDA model, which is a topic generation tool by Blei et al. [6] proposed in 2003. LDA is a three-level hierarchical Bayesian model, in which a topic is simulated as the distribution of different words, each article is constituted by a mixture of several different topics. So the topic generation process is a probabilistic generation process. LDA uses a k-dimensional Dirichlet random variable to represent the probability distribution over document and topics, simulating the generation process of documents, which is mainly used to identify the hidden topic information from large-scale document set or corpus. In recent years, LDA model and its various extended LDA models have increasingly been used in image processing, natural language processing and other fields. Moreover, in the context of tagging systems, where multiple users are bookmarking resources with multiple tags, the resulting topics can reflect a shared view of users on the document, and the tags belonging to the topics can reflect a common vocabulary. So it is possible to consider that using LDA model to extract topics information from social tagging networks.

2.2. Modularity Optimization

The modularity function Q is the most widely used indicator to characterize the strength of community features, which is first proposed by Newman et al. [11] in 2004. And with times go, it becomes a standard to measure how is the result of community detection, which is defined as:

$$Q = \sum_{i=1}^{k} (e_{ii} - a_i^2) \tag{1}$$

Where e_{ii} is the fraction of the number of edges in community i to the number of edges in the graph, and $a_i = \sum_j e_{ij}$ is the fraction of the number of the edges that connect to nodes in community i to the number of edges in the graph. And k is the final community number of community detection. The range of Q function is $[0, 1]$, and in practice it is found that a value above 0.3 is a good indicator of significant community structure in a network, and the larger the Q value is, the better the quality of the community structure in network. Therefore, the

modularity methods achieve the optimal clustering results by maximizing Q value. However, it is a NP-hard problem to maximizing Q. So a lot of heuristic approaches, which try to approximate the optimal modularity value, has been proposed [12]. Such approaches include greedy agglomeration [13], mathematical programming [14], spectral methods [15], simulated annealing [16] and so on. However, in Fortunato's work [17], they found that, modularity optimization may fail to identify communities smaller than a certain scale, which depends on the total size of the network and the degree of interconnectedness of the communities.

Besides, in real world network, some parts in communities are overlapping, that is, some nodes can belong to multiple communities with multiple community attributes. For example, a scholar may collaborate with others in many different areas in scientific collaboration networks; a user who has a broad range of interests may be associated with more than one community. Therefore, Palla et al. [18] proposed a clique percolation method (CPM), which can detect overlapping communities, but is not suitable for detecting hierarchical structures. Huang et al. [19] proposed a parameter-free algorithm SHRINK, which can not only discover overlapping and hierarchical communities but also the hub nodes and outliers among them. And based on their work, Lin et al. [20] considered that there are many incomplete information networks with missing a lot of edges common in real world networks, in which nodes are known and edges are locally known. They proposed a hierarchy clustering method used for community detection in incomplete information networks with missing edges, that is, distance-based shrinking approach (DSHRINK), which learned a distance metric from local information regions, and then estimated the distance between any pair of nodes in the network. Based on their work, a topic distance-based shrinking approach (TDSHRINK) is proposed in this paper, where the topic distance is defined by the bookmarking relationship on the same topics between users, and then the modularity based on topic distance is defined, too. Finally, clustering the nodes by using TDSHRINK algorithm to reach the goal of community detection in social tagging networks.

3. Terminology Definitions

To establish a social network graph, we must consider the interaction between users. While the interaction between two users in social tagging network can be seen as bookmark relationship on all topics. The formal definition of social tagging networks can be expressed as $TN = (User, Tag, Resource)$, where $User$ is the set of users in the networks, Tag is the set of tags, and $Resource$ is the set of resources.

Definition 1 (Topic Tagging Network): A topic tagging network is denoted as $TTN = (U, E, T, R)$, where U is the set of vertices, $E \subseteq U \times U$ is the set of edges, which represents topic distance between users. $T = (Topic_1, Topic_2, \ldots, Topic_k)$ is the set of topics that are extracted from tags, in which each topic is made up of several tags and $Topic_i = \{tag | tag \in Topic_i\}$, and R is the set of resources in tagging network.

In this paper, we extract the topics from tags in social tagging network by using LDA model, to capture the potential semantic relationships between tags and topics.

Definition 2 (Topic Distance): The topic distance on the same topic between any two users (v_i, v_j) is expressed as the bookmark relationships on the topic, which is measured by cosine similarity between them and the formula is as follows.

$$td_k(v_i, v_j) = \frac{\sum_{t \in Topic_k} [g(v_i, t) g(v_j, t)]}{\sqrt{\sum_{t \in Topic_k} g^2(v_i, t) \sum_{t \in Topic_k} g^2(v_j, t)}} \quad (2)$$

Where $v_i, v_j \in U$ and $i \neq j$. While there are several tags in a topic, $g(v_i, t)$ is the fraction of the number of resources that user v_i bookmarked with tag t to the number of the total resources that user v_i bookmarked. The topic distance is to measure the topic similarity between users, the range of which is $[0, 1]$, the smaller the value is, the higher similarity users have.

Definition 3 (Average Topic Distance): Since there are K topics in total, any pair of users have K topic distance, that is, $td_1, td_2, \ldots, td_k$. The average topic distance between any two users is calculated with Equation (3).

$$\overline{td(v_i, v_j)} = \frac{\sum_{k=1}^{K} p_k * td_k(v_i, v_j)}{K} \quad (3)$$

Where $p_1, p_2, \ldots p_k$ is the probability that each topic appeared respectively.

Definition 4 (Initial Community): Set a topic threshold α, and a initial community $IC(v_i)$ is a set of users whose average topic distance in the range of the topic threshold, which is defined as:

$$IC(v_i) = \{v_j | \forall v_i, \exists v_j, \overline{td(v_i, v_j)} \leq \alpha, v_i \in U, v_j \in U, i \neq j\} \quad (4)$$

Where α is the community radius of the initial community.

Definition 5 (Community Center): Given any three community C_i, C_j, C_k with the known community center is $cc(i)$, $cc(j)$, $cc(k)$ respectively, while the topic distance of each two communities are regarded as $d(i,j)$, $d(i,k)$, $d(j,k)$. Therefore, when C_i and C_j clustered into a new community C_m, the community center $cc(m)$ of which is determined by the two initial communities, and the topic distance between the new community C_m and C_k is calculated with Equation (5).

$$d(m,k) = \frac{1}{2}\sqrt{2d^2(i,k) + 2d^2(j,k) - d^2(i,j)} \quad (5)$$

The community center $cc(i)$ of initial community $IC(v_i)$ is the node v_i. And when clustered into new community, the new community center is determined by Definition 5, while the topic distance is calculated by Equation (5).

For convenience, we map the topic tagging network into a two-dimensional map, each node v_i has two coordinate values (x_i, y_i), and the geometrical distance between each pair of nodes (v_i, v_j) is $d(i,j) = \sqrt{(x_j - x_i)^2 - (y_j - y_i)^2}$. Therefore, corresponding to the Definition 5 in the paper, the topic distance between each pair of nodes is $\overline{td(v_i, v_j)} = \sqrt{(x_j - x_i)^2 - (y_j - y_i)^2} = d(i,j)$, which is show as Figure 1.

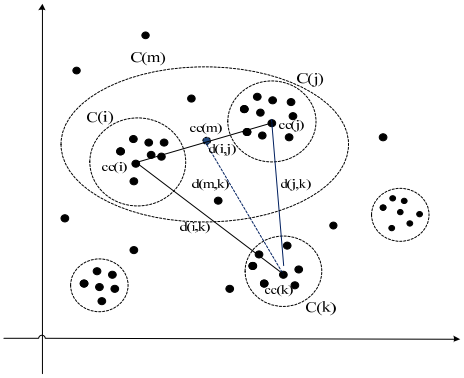


Figure 1. The Sketch Map of the Distance Calculation between Communities

4. Modularity Clustering Algorithm Based on Topic Distance

Based on the distance-based modularity and DSHRINK algorithm that Lin et al. proposed [20], we propose a topic distance-based modularity clustering algorithm TDSHRINK, which detects communities according to the topic distance between users, that is, the users with shorter topic distance will be grouped into the same community and the users with longer distance into different communities. And it can also detect overlapping communities in social tagging networks.

4.1. Topic Distance-based Modularity

Before the TDSHRINK algorithm, the topic distance-based modularity is defined as follows.

Definition 6 (Topic Distance-based Modularity): Given a topic tagging network $TTN = (U, E, T, R)$ and the communities $C = \{C_1, C_2, \ldots, C_h\}$, the topic distance-based modularity Q_{td} is defined as:

$$Q_{td} = \sum_{p=1}^{h} \left[\frac{D_p^I}{D^T} - \left(\frac{D_p^E}{D^T} \right)^2 \right] \quad (6)$$

Where h is the number of communities, $D_p^I = \sum_{v_i, v_j \in C_p} \overline{td(v_i, v_j)}$ represents the sum of average topic distance between any pair of nodes within community C_p, $D_p^E = \sum_{v_i \in C_p, v_j \in U} \overline{td(v_i, v_j)}$ represents the sum of average topic distance between any node in community C_p and any node in the topic tagging network TTN, and $D^T = \sum_{v_i, v_j \in U} \overline{td(v_i, v_j)}$ represents the sum of average topic distance between any two nodes in TTN.

As the range of modularity that Newman proposed in 2004 is $[0, 1]$ [11], in this paper, the range of topic distance-based modularity is $[0, 1]$. If $Q_{td} = 0$, it means that all the nodes are either grouped into one community or grouped into different communities randomly. And the larger value of Q_{td}, the better the quality of clustering result.

To enhance the efficiency of the algorithm, we calculate the modularity Q_{td} incrementally, that is, the gain of merging the two communities C_p and C_q into a new community, which is called topic distance-based modularity gain ΔQ_{td}. The calculational formula is as follows.

$$\Delta Q_{td} = Q_{td}^{C_p \cup C_q} - Q_{td}^{C_p} - Q_{td}^{C_q} = \frac{2D_{pq}}{D^T} - \frac{2D_p^E D_q^E}{(D^T)^2} \quad (7)$$

Where $D_{pq} = \sum_{v_i \in C_p, v_j \in C_q} \overline{td(v_i, v_j)}$ represents the sum of average topic distance between nodes in community C_p and nodes in community C_q.

The topic distance-based modularity defined above is a metric to evaluate the quality of a partition. And we use the gain of topic distance modularity to control the cluster process to get a good result of community detection.

4.2. Topic Distance-based Modularity Clustering Algorithm

The topic distance-based modularity clustering algorithm TDSHRINK is presented in Table 1. The approach can be divided into two phases. Firstly, building a topic tagging network TTN, and calculating the average topic distance $\overline{td(v_i, v_j)}$ between any pair of nodes. Secondly, (1) finding all initial communities according to Definition 4. (2) For each initial community, we find its community center by Definition 5, and calculate the topic distance between any pair of initial communities by Equation (5) to store an array. (3) find the two communities that their topic distance is smallest, and calculate the topic distance-based modularity gain ΔQ_{td}. If $\Delta Q_{td} > 0$, it means that the merger of the two communities can increase the topic distance-based modularity Q_{td}. Then merge the two communities into a new one. Otherwise, do not merge and continue to find two communities that have less smallest distance. Repeat until all clusteres are “visited” and merging the last two communities can’t increase Q_{td}. Finally, the communities are presented.

5. Experiments

In this section, we use two real-world data sets to validate the effectiveness and efficiency of the approach we proposed.

5.1. Data Sets

(1) CiteULike Dataset

The dataset in this paper is from the available datasets in CiteULike, which is a free service for managing and discovering scholarly references provided by Springer. When a user

in interested in an article, he or she can add it into his or her own library with several related tags. The format of CiteULike data includes four fields, which are article id, user name (a salted MD5 hash of the true username), the date and time and tag. And if a user posts an article with several tags, then this will result in several rows, which is shown in Table 2.

Table 1. The Description of TDSHRINK

Algorithm 1 **TDSHRINK**
Input: $TN = (User, Tag, Resource)$, topic threshold α;
Output: $C = \{C_1, C_2, \ldots, C_i, \ldots, _n\}$:
Process:
1 Initialize, build the topic tagging network $TTN = (U, E, T, R)$;
for each $v_i \in U$ **do**
  for each $v_j \in U \wedge v_i \neq v_j$ **do**
    Calculate $\overline{td(v_i, v_j)}$ according to Equation 2;
    $D^T += \overline{td(v_i, v_j)}$;
  end
end
for each $v_i \in U$ **do**
  if $v_i.visited$ **then continue**
  Find a intial community $IC(v_i)$ according to Definition 4;
  for each $v_j \in IC(v_i)$ **do**
    $v_i.visited = true$;
  end
  $C \leftarrow C \cup IC(v_i)$;
end
for each $C_i \in C$ do
  for each $C_j \in C \wedge i \neq j$ **do**
    Calculate the distance $d(i, j)$ between C_i and C_j according to Equation 5 and store the distance into $dvalue$ array;
  end
end
While true do
  Select the smallest distance $d(p, q)$ in $dvalue$ array;
  Calculate corresponding ΔQ_{td} according to Equation 7;
  if $\Delta Q_{td} > 0$ **do**
    $C_m \leftarrow C_p \cup C_q$ and update $dvalue$ array;
  else
    $d(p, q) = \infty$;
  if $\mathbf{MIN}\, d(i, j) \geq \infty$ **then break**;
end
return C;

Table 2. The Format of the Raw Data

Article id	Username(MD5)	Date and time	tag
9168221	654442b4eaff2791d205c4abdeb99375	2012-01-01 00:21:27.814194+00	pvalue
5827136	654442b4eaff2791d205c4abdeb99375	2012-01-01 00:22:17.990863+00	pvalue
10186672	aac984847268804c15d115fbee0b3652	2012-01-01 00:26:26.822489+00	rsvp_iconchat
10186790	9730960ede281beae7419006b47dbf41	2012-01-01 01:55:47.960338+00	motivation
10186791	9730960ede281beae7419006b47dbf41	2012-01-01 01:58:43.275636+00	massively_multiplayer_online_games

Since the raw data is large, to facilitate the latter experiments, we intercept all the data from Jan. 2012 to Dec. 2012 as a data set. In addition, we focus on the study of tags that user used and the corresponding article resources, then calculate the topic distance between users, which is not directly related to the time. Therefore, we need to extract three fields of data from

the raw data, which are article id, username and tag respectively. In data cleaning, we firstly apply stemming to the tags, split the tags that are linked by symlinks into a few words. In addition, delete those insignificant tags, such as, “no-tag”, prepositions, pure digital tags and so on. Finally, in order to simplify the subsequent calculations and ensure the accuracy rate, we delete the tags that are bookmarked by less than 10 users. After data cleaning, there are 18512 tags, 13086 users, 137306 articles. Through the LDA topic extraction procedure that Zhou Li publicly available, we get 100 topics ultimately and each topic includes several tags with corresponding probability, which is shown in Table 3.

Table 3. The Data of Part Topic with Corresponding Tags and Probability

Topic1	p1	Topic 2	p2	Topic 3	p3
paper	0.0869874	healthcare	0.0514422	attention	0.0245588
lanlsec	0.0159081	privacy	0.0332143	disorder	0.0194616
holopedia	0.0144484	security	0.0276437	auditory	0.0160443
access	0.00988849	mhealth	0.0222972	deficit	0.0156371
open	0.00894257	rechtslinguistik	0.0216415	adhd	0.015277

(2) DBLP-A dataset

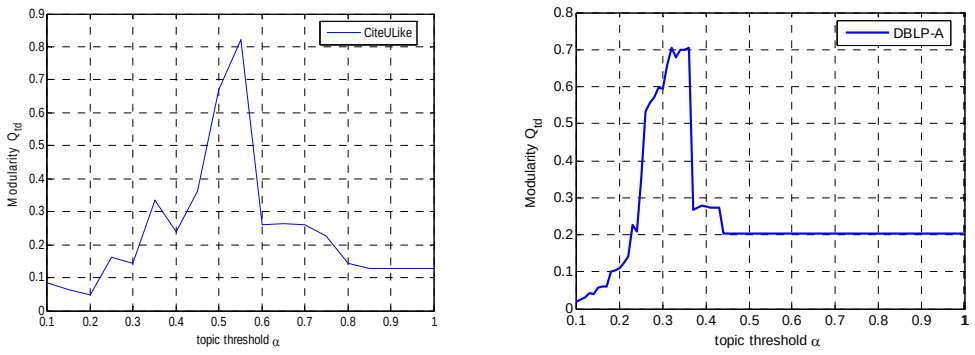
DBLP-A is the data set extracted from DBLP website which provides bibliographic information on computer science journals and proceeding. In order to compare with the algorithm proposed in paper [20], we process the data as the way Lin et al. do. To fit the approach we proposed, we carry out some processing of the data to build topic tagging network, view the articles that authors coauthored as resources of TTN. And the choice of tags is important, to ensure the relevance of topics, we view the words in the title as tags, then apply the standard text processing, such as stemming, stop words removal. The rest of the process is similar to the above section of CiteULike. The processed data is shown in Table 4.

Table 4. The Data Sets used in Experiment

Dataset	Users	Tags	Resources	Topics
CiteULike	13086	18512	137306	100
DBLP-A	5417	3393	5455	6

5.2. The Choice of Topic Threshold α

From the above analysis, the larger the value of topic distance-based modularity Q_{td}, the better the clustering results are. Moreover, the choice of topic threshold will influence the formation of initial communities, and further influence the effectiveness of results. Therefore, in this paper, the topic threshold α is determined by Q_{td}, that is, the value that makes Q_{td} of initial communities largest, as the topic threshold of our approach. The range of α is $[0,1]$, in steps of 0.01. To generate initial communities according to corresponding α value, and calculate corresponding Q_{td}. The result is shown in Figure 2.



(a)CiteULike (b)DBLP-A

Figure 2. The Choice of Topic Threshold α

We can obtain that the topic distance-based modularity Q_{td} of the two datasets first rise to peak, then decrease, and ultimately keep stable. It is mainly because that the smaller the threshold is, the more the number of initial communities, and the more scatter the users in network. While as the increasing of threshold, the number of initial communities becomes small, the users become concentrated. At last, the modularity Q_{td} become invariant when α reaches a certain value. While the corresponding topic threshold α of largest Q_{td} value are 0.55 and 0.32 respectively, which are as the value of topic threshold α in our approach.

5.3. Evaluation Measures

In order to measure the effectiveness of our approach and compare with the approach that Lin et al. proposed [20], we adopt the same evaluation criterion, that is, Purity, to evaluate the quality of the communities generated by different approaches. The definition of purity is as follows: each cluster in first assigned with the most frequent class in the cluster, and then the purity is measured by computing the number of the instances assigned with the same labels in all clusters, which is calculated with Equation (8).

$$Purity = \frac{1}{n}\sum_{i=1}^{k} max_j \left| C_i \bigcap l_j \right| \tag{8}$$

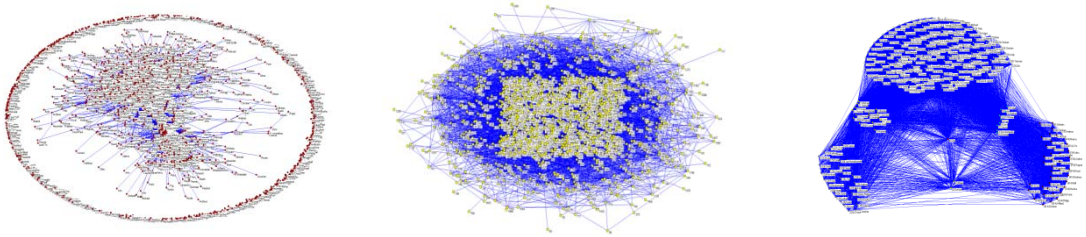
Where $\{C_1, \ldots, C_k\}$ is the set of clusters that generated by each detection approach, l_j is the j-th class label. The range of purity is $[0, 1]$, and the higher purity means the higher accuracy of the method. The community structure generated by each compared method will be evaluated using the true label of each node. Since each user can have multiple interests in CiteULike and each author can have multiple research areas in DBLP as its class labels, that is, each node can belong to overlapping communities. Therefore, we compute the purity of the clustering results based on label separately, and the average results over 100 or 6 labels are reported.

5.4. Experiment Results and Analysis

To verify the availability and effectiveness of the TDSHRINK algorithm we proposed, we have experiment on the true dataset from the social tagging network CiteULike. And at the same time, to compare with the DSHRINK algorithm that Lin et al. proposed [20], we experiment with the same dataset DBLP-A.

5.4.1. Visualization and Analysis of Results

As a visualization tool used usually in complex network, Pajek can effectively analyze and demonstrate the structural properties of complex networks. In this paper, we choose Pajek to demonstrate the effect of results of community detection.



(a) Topology of original network (b) Distribution of initial communities (c) Display of part final results

Figure 3. The Process Display from Original Network to Final Results of DBLP-A

Figure 3(a) is the original network of DBLP-A dataset we constructed, where each node represents one of the authors, the blue line represents the average topic distance between any pair of users. We can see from the figure, nodes in the network are divided into two parts, where nodes inside are closely connected, outside are not connected with edges, which indicates that the nodes outside are isolated nodes, and corresponding to the result in Table 4. Figure 3(b) is

display of results of initial communities generated by the algorithm we proposed. Different from Figure 3(a), in this figure, each node represents an initial community, and there are 1634 initial communities in total. We can find that those initial communities are closely linked, but the community structure is not very obvious, and further detected can get final result of community detection. We only select three representative communities for display in this paper, which are community numbered by 23, 129, and 614 respectively, as shown in Figure 3(c). Each node in the figure represents a user, each dense cluster represents a community, and the points that linked between communities are nodes that belong to overlapping communities.

Table 4. Results of Community Detection

	CiteULike	DBLP-A
Number of nodes	13086	5417
Clustering coefficient	0.35187	0.75708
Number of initial communities	4825	1634
Number of final communities	2922	1033
Number of isolated nodes	118	1003
Average size of community	94	16
Number of nodes in maximum community	1570	289

Through statistics in Table 4, we can find that the proposed algorithm can effectively detect communities on two kind of networks, the number of community, the average size of community is reasonable. There is a term named number of isolated nodes, which are some nodes that do not participate in the community detection. And it indicates that there are some users don't involve in the bookmark of topics, in other words, the tags that they used are not included in topics, part of users are interested in certain topic, other are not, which is reasonable. Clustering coefficient is a parameter to measure network community effect in complex network, the range of which is $[0, 1]$. In actual network, clustering coefficient is much smaller than 1, but much greater than $O(\frac{1}{n})$. Therefore, the average clustering coefficient of CiteULike is 0.35187, which is living up to actual network and DBLP-A is 0.75708, which indicates that DBLP-A has high network clustering effect, closely connected between nodes.

On the other hand, we can get obviously from Table 4, there are less isolated nodes in CiteULike dataset than DBLP-A dataset, which is mainly because the application background of this algorithm is tagging network, while DBLP-A is coauthor network without tag information. Therefore, there are more isolated nodes of result in DBLP-A, but the remaining nodes that participated in community detection have high clustering effect.

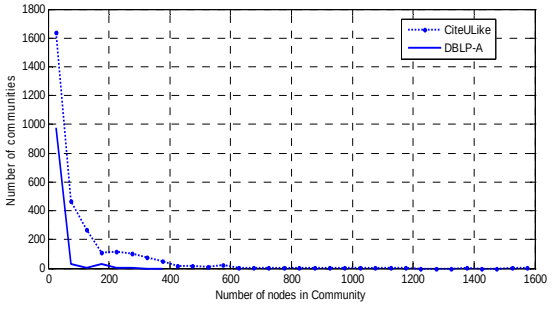


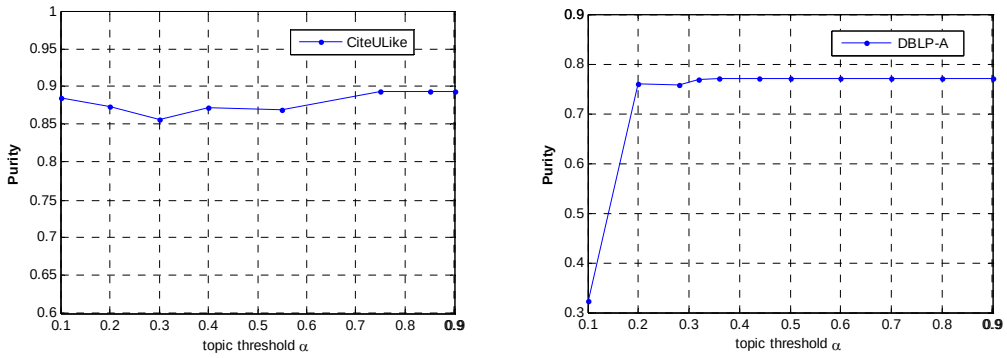
Figure 4. The Distribution of Final Communities of Two Datasets

We have made statistics of distribution of number of similar community size of two datasets CiteULike and DBLP-A respectively, which is shown in Figure 4. We can find that the community results of two datasets are mostly concentrated on small number of nodes, and there are most communities in $[0, 200]$ range, which indicates that the network clustering effect is strong, and the results of community detection are satisfactory. And the distribution of communities of CiteULike is more comprehensive, where large communities that number of which is in $[1000, 1600]$ range, middle communities in $[200, 1000]$ range, small communities in

$[0, 200]$ range are detected, which also indicates that users have a wide range of interest on multiple topics.

5.4.2. Effectiveness and Comparison Analysis

As the application background of the TDSHRINK algorithm we proposed is different from DSHRINK algorithm that Lin et al. proposed [20], this algorithm is applied to tagging network, such as CiteULike, Del.icio.us and so on. While DSHRINK algorithm is applied to incomplete information network. Therefore, we make statistic about Purity value of clustering results of TDSHRINK algorithm in two datasets CiteULike and DBLP-A in different topic thresholds α, which is shown in Figure 5.



(a) CiteULike (b) DBLP-A

Figure 5. Accuracy of TDSHRINK Algorithm in Different Topic Thresholds α

From Figure 5(a) we can find that in CiteULike dataset the change of topic threshold α produces little influence on Purity value, which is in $[0.8, 0.9]$ range, indicating that TDSHRINK algorithm we proposed is effective in social tagging networks, and has high accuracy rate in total. In Figure 5(b), the change of topic threshold will cause dramatic changes on Purity value, when α is small, such as 0.1, Purity value is only 0.323, while α increases 0.2, Purity value jumps to 0.759. And the topic threshold α we chose in the paper is 0.32, corresponding Purity value is 0.772, which is equivalent with the overall accuracy of Lin's paper [20] in the same dataset DBLP-A, indicating that the proposed algorithm is effective. Moreover, when the topic threshold α is chosen appropriately, the accuracy of TDSHRINK algorithm we proposed is higher than Lin's algorithm. In addition, from Figure 5(a) to Figure 5(b), we can clearly find that the Purity value of TDSHRINK algorithm in CiteULike dataset is higher than DBLP-A dataset, which is mainly because the application background of TDSHRINK algorithm is social tagging networks.

And that, the number of final communities of CiteULike and DBLP-A in different topic threshold α is shown in Figure 6.

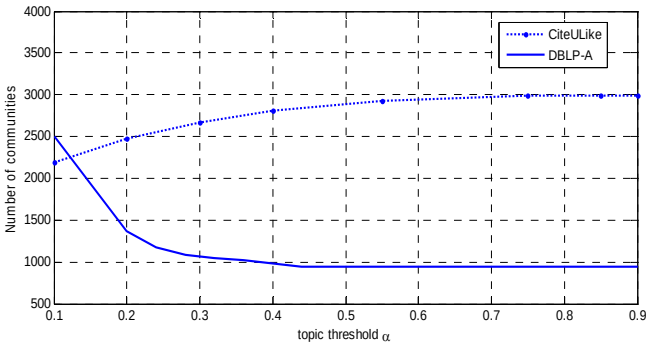


Figure 6. The Distribution of Number of Final Communities in Different Topic Threshold α

As shown in Figure 6, there is large difference on the results of two datasets, the number of communities in CiteULike is concentrated in $[2000, 3000]$ range, while DBLP-A is in $[950, 2500]$ range, which is mainly because the data size of CiteULike is about three times of DBLP-A, and users in CiteULike are linked closely, so the number of communities is concentrated. On the other hand, we find that the number of communities has large difference of two datasets on the same topic threshold, which also indicates that the network structure of network can impact on the results of community detection, but also verifies the wide applicability of the algorithm from the side. At last, combining Figure 5 with Figure 6, we can see that the less number of final communities dose not correspond to a higher Purity value, which also indirectly proves that the higher topic threshold does not mean better, an appropriate topic threshold can correspond to a good result of community detection.

5.4.3. Efficiency Results

The computing time of the algorithm we proposed is mainly divided two parts, one is the calculation of topic distance between any pair of users, and the other part is the execution time of the algorithm. The calculation of topic distance can be accomplished in advance before clustering process. Therefore, the time complexity of the algorithm is mainly concentrated on the clustering process. From the process of the algorithm shown in Table 1, the time complexity of TDSHRINK algorithm we proposed is $O(m^2 \log m)$, where m is the number of initial communities. While from the description of DSHRINK algorithm Lin et al. proposed [20], the time complexity of the algorithm is $O(kn^2)$, where k is the modularity optimization times and n is the number of users in the network. Obviously, the number of initial communities is smaller than the number of users, so the time complexity of TDSHRINK algorithm we proposed is lower. And the efficiency of our algorithm is closely related to the number of initial communities, which is determined by topic thresholdα. When we choose an appropriate α, the run time of TDSHRINK algorithm we proposed is slightly less than DSHRINK, which can be reflected in Figure 7.

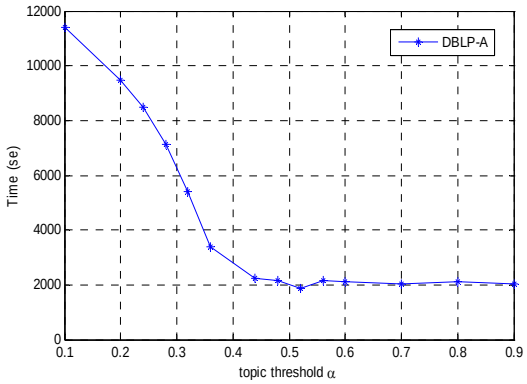


Figure 7. The Run Time of TDSHRINK Algorithm with the Different Topic Threshold α

6. Conclusion

In this paper, we presented a new clustering algorithm which considered topic information in social tagging networks. Different from the traditional topology-based community detection approaches, we start from the viewpoint that users that interested in same topics belong to the same community, add topic information into community detection, use LDA topic model to extract topics hidden in tags, and consider that a user may be interested in multiple topics to detect overlapping communities. Therefore, we define a topic distance between any pair of users in this paper to cluster users, and further propose a topic distance-based modularity function on the basis of Lin et al. [20]. Based on those, we propose a modularity clustering algorithm based on topic distance in social tagging networks, which is called TDSHRINK. Experimental results on real dataset CiteULike show that the algorithm can efficiently detect community structures in social tagging networks, and results on the same

dataset DBLP-A that Lin et al. used [20] show that, our approach is improved in accuracy and efficiency to some extent. There are several interesting directions for future work. We only use LDA model to extract topics, and don't evaluate the quality of the extracted topics, which requires further analysis. In addition, the proposed approach is mainly used in tagging network, the next step will be considered to apply to other types of networks, such as micro-blogging.

Acknowledgements

This paper is supported by the following foundation or programs, including Natural Science Foundation of Chongqing of China (cstc2012jjA40027); Youth Scientific Research Project of Chongqing University of Posts and Telecommunications of China (A2012-87); National Social Science Fund Project (13CGL146); Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ130518).

References

[1] Ding Y. Community detection: Topological vs. topical. *Journal of Informetrics*. 2011; 5(4): 498–514.
[2] Girvan M, Newman MEJ. *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences. 2002; 99(12): 7821-7826.
[3] Kernighan B W, Lin S. An efficient heuristic procedure for partitioning graphs. *Bell Systems Technical Journal.* 1970; 49(2): 291-307.
[4] Pothen A, Simon HD, Liou KP. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal of Mathematic Analysis and Applications.* 1990; 11(3): 430–452.
[5] Zhang H, Qiu B, Giles CL, et al. *An LDA-based community structure discovery approach for large-scale social networks.* IEEE Conference on Intelligence and Security Informatics. New Jersey. 2007: 200-207.
[6] Blei DM, Ng A Y, Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 2003; 3: 993-1022.
[7] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. *The author-topic model for authors and documents*. Proceedings of the 20th conference on Uncertainty in artificial intelligence. Virginia. 2004: 487-494.
[8] Zhou D, Manavoglu E, Li J, et al. *Probabilistic models for discovering e-communities.* Proceedings of the 15th International Conference on World Wide Web. NY. 2006: 173-182.
[9] Sachan M, Contractor D, Faruquie TA, et al. *Using Content and Interactions for Discovering Communities in Social Networks*. Proceedings of the 21st international conference on World Wide Web. NY. 2012: 331-340.
[10] Griffiths TL, Steyvers M. Finding Scientific Topics. *Proceedings of National Academy of Sciences of the United States of America.* 2004; 101(Suppl 1): 5228-5235.
[11] Newman MEJ, Girvan M. Finding and evaluating community structure in networks. *Physical Review E.* 2004; 69 (2): 026113.
[12] Good BH, de Montjoye YA, Clauset A. Performance of modularity maximization in practical contexts. *Physical Review E*. 2010; 81(4): 046106.
[13] Wakita K, Tsurumi T. *Finding community structure in mega-scale social networks.* Proceedings of the 16th international conference on World Wide Web. NY. 2007: 1275-1276.
[14] Agarwal G, Kempe D. Modularity-maximizing graph communities via mathematical programming. *The European Physical Journal B.* 2008; 66(3): 409-418.
[15] Shiga M, Takigawa I, Mamitsuka H. *A spectral clustering approach to optimally combining numerical vectors with a modular network.* Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. NY, 2007: 647–656.
[16] Guimera R, Amaral LAN. Functional cartography of complex metabolic networks. *Nature*. 2005; 433(7028): 895–900.
[17] Fortunato S, Barthelemy M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences.* 2007; 104(1): 36–41.
[18] Palla G, Derenyi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. *Nature.* 2005; 435(7043): 814-818.
[19] Huang J, Sun H, Han J, et al. *SHRINK: a structural clustering algorithm for detecting hierarchical communities in networks*. Proceedings of the 19th ACM international conference on Information and knowledge management. NY. 2010: 219–228.
[20] Lin W, Kong X, Yu PS, et al. *Community Detection in Incomplete Information Networks*. Proceeding of the 21st international conference on World Wide Web. NY. 2012: 341-350.