Knowledge-Based Systems 26 (2012) 164–173

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Topic oriented community detection through social objects and link analysis in social networks

Zhongying Zhao [a,b,c,*], Shengzhong Feng [b], Qiang Wang [b], Joshua Zhexue Huang [b], Graham J. Williams [b], Jianping Fan [b]

[a] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[c] Graduate School of Chinese Academy of Sciences, Beijing 100080, China

ARTICLE INFO

Article history:
Received 23 June 2010
Received in revised form 28 July 2011
Accepted 28 July 2011
Available online 4 August 2011

Keywords:
Social networks
Community detection
Link analysis
Social objects clustering

ABSTRACT

Community detection is an important issue in social network analysis. Most existing methods detect communities through analyzing the linkage of the network. The drawback is that each community identified by those methods can only reflect the strength of connections, but it cannot reflect the semantics such as the interesting topics shared by people. To address this problem, we propose a topic oriented community detection approach which combines both social objects clustering and link analysis. We first use a subspace clustering algorithm to group all the social objects into topics. Then we divide the members that are involved in those social objects into topical clusters, each corresponding to a distinct topic. In order to differentiate the strength of connections, we perform a link analysis on each topical cluster to detect the topical communities. Experiments on real data sets have shown that our approach was able to identify more meaningful communities. The quantitative evaluation indicated that our approach can achieve a better performance when the topics are at least as important as the links to the analysis.

1. Introduction

With social networks becoming popular (such as Flickr, YouTube, LiveJournal, Facebook, Digg, MySpace, DBLP collaboration network, etc.), analyzing such network data has become an increasingly important research issue. Community detection, as a major topic in social network analysis, has received a great deal of attention [1–3]. Discovering inherent community structures can help us understand the networks more deeply and reveal interesting properties shared by the members. People belonging to the same community are more likely to have common hobbies, social functions, occupations, interests on some topics, viewpoints etc. Therefore, the identified communities can be used in collaborative recommendation [4], information spreading [5], knowledge sharing [6], and other applications, which can benefit us greatly.

Existing studies on community detection mainly focus on link analysis or topological structure of the network [7–14]. Communities identified by those works often incorporate different topics since stronger connections represent the interactions that occur across several different topics, which confuses the meanings of the community. Look at the example illustrated in Fig. 1. Fig. 1(a) is a social network consisting 9 nodes and 11 edges. The nodes represent the members involved in the social activities and the edges represent the social relations of interactions or communications. The weight attached to each edge represents the strength of connections between the corresponding members. In addition, we assume the topics of each member have been extracted from social objects through clustering, and the topics are labeled at each node, such as football, music or both. Fig. 1(b) shows the result of discovered communities based on link analysis. We can see that members within a community are connected, but they have different topics of interest. In the left community of Fig. 1(b), there are 4 members interested in 'football' and other 3 members interested in 'music'. This shows that the results from link analysis have ambiguous meanings of communities.

Social objects like emails or blogs often imply the topics that are shared by people. This has motivated research on community discovery through analyzing the contents of the social objects [15–17]. Each community identified by this kind of method often has one common topic. However, such community often contains weakly connected people since it is common that some people are often not connected, especially in distributed environments, i.e., they do not know each other and never communicate. Fig. 1(c) shows the results from clustering of social objects on the network of Fig. 1(a). We can see that members within a community have a

* Corresponding author at: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.
E-mail addresses: zy.zhao@siat.ac.cn, zy.zhao10@gmail.com (Z. Zhao), sz.feng@siat.ac.cn (S. Feng), qiang.wang1@siat.ac.cn (Q. Wang), zx.huang@siat.ac.cn (J.Z. Huang), Graham.Williams@togaware.com (G.J. Williams), jp.fan@siat.ac.cn (J. Fan).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.07.017

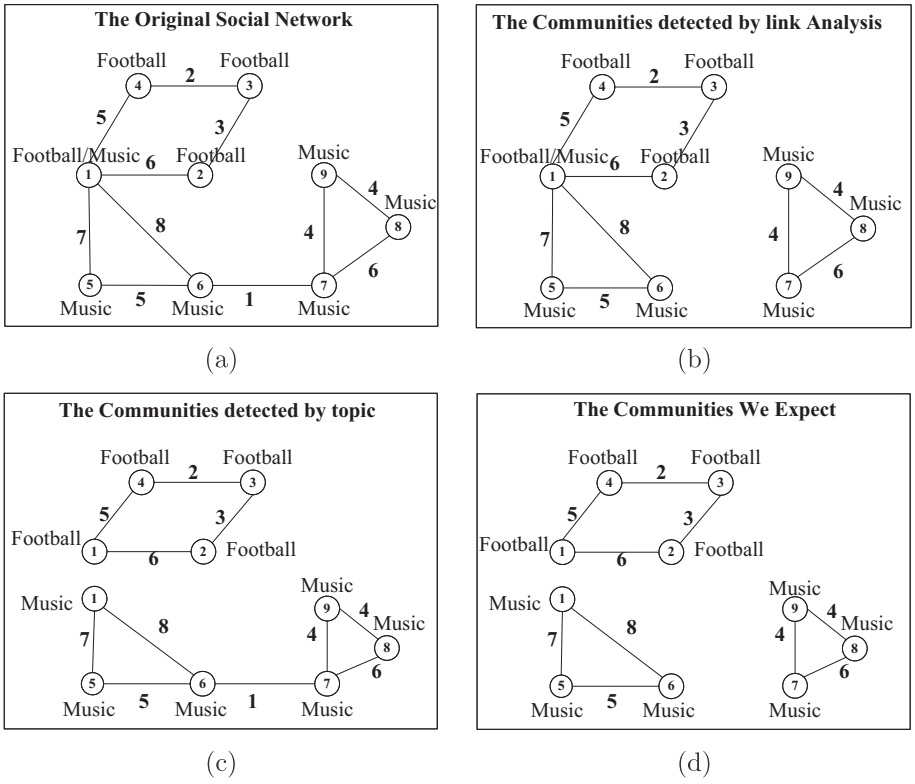


Fig. 1. An example to illustration the motivation of our work.

common topic, but they are not closely connected. In other words, the resulting communities formed on topics only can not reflect the strength of social relations.

To sum up, neither link analysis nor social objects clustering alone is sufficient in determining meaningful communities. Fig. 1(d) shows an ideal result. The members within one community are closely connected, meanwhile they have the same interested topics. This is the result we aim to achieve in this paper.

In this paper, we propose a topic oriented community detection approach which combines social objects clustering and link analysis. Firstly, all social objects are clustered into different topics. Then, the members involved in those social objects are divided into different topical clusters, each corresponding to a certain topic. Finally, the link-based community detection for each topical cluster is performed to differentiate the strength of connections between members.

Compared to the existing work, our approach can identify communities from the perspective of both topics and the link structures. From this result, we can easily find which people are attracted to which community, and by what topic. Such findings can be used to improve the performance of direct marketing and collaborative recommendations.

The rest of the paper is organized as follows. Section 2 reviews the related work. In Section 3, we present the topic oriented community detection approach based on social objects clustering and link analysis. In order to verify our approach, we conducted extensive experiments on real life data sets. The experimental design and results analysis are given in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

A wide range of works have been done to discover communities in a network [7–14]. According to the community detection strategy, the previous works can be classified into optimization based methods and heuristic methods. The optimization based methods include spectral methods and local search based methods. Spectral methods aim to minimize the defined cut-function (e.g., [7]), while local search based methods aim to optimize an objective function such as the function of 'modularity' (e.g., [8,18–23]). The modularity function is used to evaluate the quality of a particular division of a network into communities [24,25]. The heuristic methods often design a graph clustering algorithm based on intuitive assumptions (e.g., MFC algorithm [9], HITS algorithm [10], GN algortihm [11], WH algorithm [12], CPM algorithm [13], FEC algorithm [14]).

Those works have gained success in some applications but they mainly focus on the topological structures or linkage patterns of networks, ignoring the interested topics shared by members. As a result, a community often contains members interested in different topics, which misleads or mixes the meanings of the community.

Another related work is topic modeling through analyzing the contents of social objects. There are several topic models, such as pLSI [26], LDA [27], AT [28], etc. Interactive applications or social network analysis motivate research on topic modeling or topic-based community detection. Zeng et al. [15] proposed a framework for analysis of user activity on an interactive website. User activity analysis tasks, such as user group discovery, can be performed in the framework. McCallum et al. [16] presented the Author-Recipient-Topic model to discover the discussion topics of social networks. Tian et al. [17] proposed OLAP-style aggregation strategies to partition the graph according to attribute similarity, so that nodes within one community share the same attribute values.

The above methods aim to group members interested in common topics into one community. However, they ignore the relationships between members. As shown in Fig. 1(c), the generated communities tend to have very low connectivity.

In fact, both the topological structure and social objects associated with members are important for community detection. The existing approaches mentioned above consider only one aspect but ignore the other. As a result, the identified communities either

contain more than one topic of interest shared by members, or have loose connections between members within the same group. To address this problem, we propose here an approach that bring these two bodies of work together.

3. Topic oriented community detection approach

In this section, we present a topic oriented community detection approach based on social objects clustering and link analysis. This approach can identify the topical communities which reflect the topics and strength of connections simultaneously. We first give the framework of our method in Section 3.1. We describe the social network data modeling in Section 3.2. Social objects clustering and link analysis are presented in Section 3.3 and Section 3.4. We summarize our approach in Section 3.5.

3.1. Framework

The framework for topic oriented community detection is illustrated in Fig. 2. The whole process for identifying communities is divided into 4 key modules.

(1) Social network data modeling: This module aims to structure the datasets into formal models for processing.
(2) Social Objects Clustering: According to the contents of the social objects, we cluster them into different clusters. Each cluster represents a topic shared by the members of the cluster.
(3) Social Members Partitioning: Based on the associations between members and social objects, we partition the members involved in social objects into different topical clusters.
(4) Link Analysis: The members in each topical cluster are connected with different strengths. We perform link analysis based community detection on each topical cluster. The communities detected in each topical cluster are regarded as topical communities.

3.2. Social network data modeling

Taking the social objects into consideration, we propose a formal graph model to describe the social networks.

Definition 1 (*EG model*). The extended graph (*EG*) is defined as a three-tuple: $EG = (U, O, E)$, where

(1) U is the set of users/members involved in the social activities. We use a circle to denote each member.
(2) O is the set of objects (or contents) that members communicate. Each object is represented by a square.
(3) E is the set of edges, representing relations that exist in *EG*. $E = E_{UO} \cup E_{UU}$, where $E_{UO} = \{(i,j)|i \in U, j \in O\}$, $E_{UU} = \{(i,k)|i \in U, k \in U\}$.

The social objects that members communicate can be classified into two kinds of situations: (1) attached to multi-members; (2) attached to only one member.

In the first situation, we consider that the edges between members are constructed because of the social objects. An example is the coauthor network whose *EG* model is shown in Fig. 3(a). In this network, each paper (social object) is attached to multi-authors. The authors (members) are connected with each other due to their collaboration on the same paper.

In the second situation, we consider the social objects to be the attributes of members, since each object is attached to only one member. An example is the blogs citation network whose *EG* model is shown in Fig. 1(a). In this network, blogs (social members) cites each other. Furthermore, each blog is associated with a text content (social object) which can be considered as the attribute of the corresponding blog.

In this paper, we consider the number of communications happened between users and the text contents of social objects. We do not concern how much effort users/members have spent since it is difficult to quantify. Therefore, we did not devise any weights for different objects in the social network modeling process. In other words, no matter what user i sends to user j, we consider i finishes one communication with j. For the coauthor network and the blogs citation network, we can see that the relations between users/members in the bottom layer of Fig. 3(a) and Fig. 3(b). Each of the relations is attached to a weight which represents the number of communications.

3.3. Social objects clustering and members partitioning

Generally speaking, people carry out social activities on social objects, such as emails and blogs. Those objects often imply the topic that people are interested in. In this paper, we only consider text social objects.

A text object is often viewed as a set of pairs $\langle t_i, mea_i \rangle$, where t_i is a term or word, and mea_i is a measure of t_i. The total number of unique terms in a text data set represents the number of dimensions. We adopt the vector space model (VSM) to represent each text object j.

$$V_j = ((t_1, mea_1), (t_2, mea_2), \ldots, (t_n, mea_n)), \quad j = 1, 2, \ldots, m, \qquad (1)$$

where

- t_i refers to the term or word,
- mea_i corresponds to the measure of t_i. In this paper, we have adopted the measure of *tf*/*idf*.
- n is the number of terms or words after preprocessing, representing the dimension of the text set.
- m refers to the number of text objects.

Then, all the text objects are denoted by matrix M_{mn}. Each row of the matrix represents a text object, while each column represents a dimension.

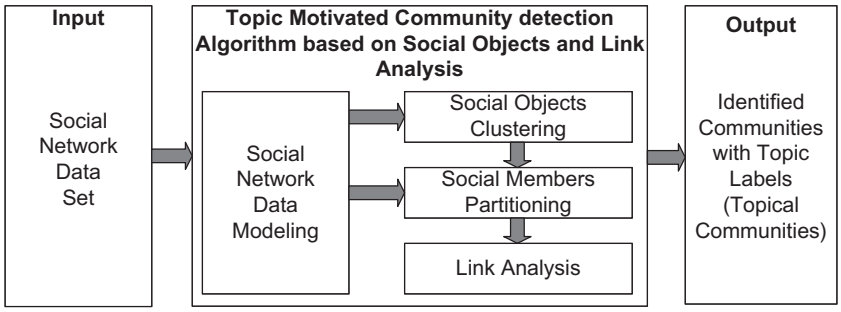
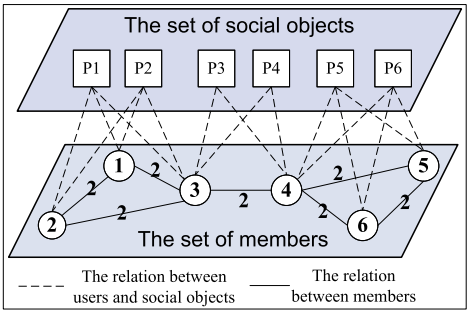
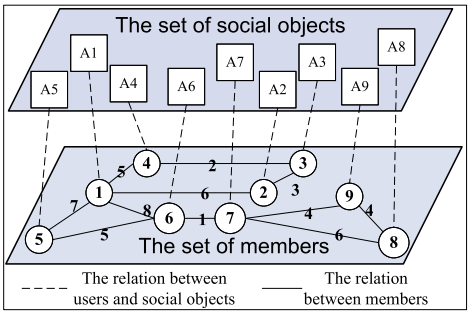


Fig. 2. The framework for detecting communities based on social objects clustering and link analysis.



(a) The EG model of a coauthor network, where $U = \{1,2,3,4,5,6\}$, $O = \{P1, P2, P3, P4, P5, P6\}$. The edge between member (author) i and object (paper) j represents the writing relation, while the edge between two members (author i and k) represents the coauthoring relation.



(b) The EG model of a blogs citation network, where $U = \{1,2,3,4,5,6,7,8,9\}, O = \{A1, A2, A3, A4, A5, A6, A7, A8, A9\}$. The edge between member i and Aj means that Aj is the attribute of member i. The edge between two members (i and k) represents the citation relation of blogs.

Fig. 3. The Extended Graph (EG) models for a coauthor network and a blogs citation network.

We assume that different dimensions have different contributions to the identification of objects in a cluster. From this point, we adopt the Entropy Weighting K-Means (EWKM) algorithm [29] to cluster the texts. The EWKM algorithm, as an extension of the k-means method, aims to calculate a weight for each dimension in each cluster. Then the weights values are used to identify the subsets of important dimensions that categorize different clusters. This is achieved by including the weight entropy in the objective function that is minimized in the k-means clustering process. We give a simple description of EWKM in Algorithm 1. For details about this algorithm, please refer to [29].

Algorithm 1. EWKM-Entropy Weighting K-Means algorithm

Input: matrix M_{mn}, the number of clusters K and parameter γ;
Output: clusters of social objects, the weights of dimensions for each cluster;
1: randomly choose K cluster centers and set all initial weights to $1/m$;
2: **while** (the objective function not getting its local minimum value) **do**
3: update the partition matrix;
4: update the cluster centers;
5: update the dimension weights;
6: **end while**;
7: **return** clusters and the corresponding dimension weights;

With the EWKM algorithm, we obtain the clusters of social objects and the dimension weights. Several top dimensions are selected to label each cluster. The result is object clusters in terms of their topics.

Given the social object clusters with labels, we partition the members involved in social objects into different topical clusters. We allow a member belonging to several topical clusters since it is common that people are interested in several different topics.

3.4. Topical community detection (or link analysis) for each topical cluster

The members in each topical cluster are often connected to each other with different strengths. Some members may communicate to each other with higher frequency, which leads to stronger connections. Others may have few or no communication which results in weak or no connection. In order to identify the tightly connected members, we have to make a link analysis on each topical cluster. We regard this process as topical community detection or link analysis.

In this process, many community detection methods can be employed, such as degree centrality, edge betweenness, random walk method, GN, cliques and so on. In this paper, we use the modularity maximization method [30] to do link analysis. This method can measure the goodness of divisions, since it is based on the greedy optimization of the quantity known as modularity [31]. Furthermore, it can be extended to large graphs.

Suppose we have divided the topical cluster $Cluster^p$ into several communities. The modularity Q^p for this cluster is defined as follows:

$$Q^p = 1/(2m)\sum_{vw}(A_{vw} - (k_v k_w)/(2m))\sum_i \delta(c_v, i)\delta(c_w, i) = \sum_i (e_{ii} - a_i^2), \quad (2)$$

where

- v, w are vertices within this topical cluster;
- A_{vw} is an element of the adjacency matrix corresponding to the topical cluster p. It represents the number of communications that have occurred between users v and w;
- $m = 1/2\sum_{vw} A_{vw}$;
- $k_v = \sum_u A_{vu}$, where u is a vertex;
- c_v is the community to which vertex v is assigned;
- i represents the community i;
- $\delta(x,y)$ is 1 if $x = y$ and 0 otherwise;
- $e_{ij} = 1/(2m)\sum_{vw} A_{vw}\delta(c_v, i)\delta(c_w, j)$;
- $a_i = 1/(2m)\sum_v k_v \delta(c_v, i)$;

The traditional modularity method starts off with each vertex being a community which contains only one member. Then it computes the changes of Q^p, chooses the largest of them, and performs the merge of communities. In order to improve the efficiency, the method in [31] maintains and updates a matrix of value of ΔQ_{ij}^p. Three data structures are maintained: (1) a sparse matrix which contains ΔQ_{ij}^p for each pair i, j of communities with at least one edge between them; (2) a max-heap H which contains the largest element of each row of the matrix ΔQ_{ij}^p along with the labels i, j of

the corresponding pair of communities; (3) an ordinary vector array with elements a_i.

The initialization for ΔQ_{ij}^p and a_i is as follows:

$$\Delta Q_{ij}^p = \begin{cases} 1/(2m)A_{ij} - (k_i k_j)/(2m) & \text{if } i,j \text{ are connected,} \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

$$a_i = k_i/(2m), \quad (4)$$

The updating rules for matrix of ΔQ^p are as follows.

$$\Delta Q_{ij}^{p'} = \begin{cases} \Delta Q_{ik}^p + \Delta Q_{jk}^p & \text{if community } k \text{ is connected to both } i,j, \\ \Delta Q_{ik}^p - 2a_j a_k & \text{if community } k \text{ is connected to } i \text{ but not to } j, \\ \Delta Q_{jk}^p - 2a_i a_k & \text{if community } k \text{ is connected to } j \text{ but not to } i. \end{cases} \quad (5)$$

The modularity maximization method we used for each topical cluster is simply described in Algorithm 2. For details, please refer to [31].

Algorithm 2. Modularity maximization algorithm

Input: topical cluster $Cluster^p$, $p = 1,2,3,\ldots$
Output: communities identified
1: calculate the initial values of $\Delta Q_{ij}^p, a_i$ according to Eq. 3, 4;
2: **while** (the current number of communities > 1) **do**
3: **for** each row of the matrix of ΔQ^p **do**
4: populate the max-heap H with the largest element;
5: **end for**
6: select the largest ΔQ_{ij}^p from H;
7: join the corresponding communities;
8: update the matrix of ΔQ^p;
9: **end while**
10: **return** communities identified;

3.5. Summary and analysis of our approach

In this section, we first summarize our approach with an algorithm. We then explain why we adopt these three main steps in this section. Algorithm 3 describes our approach as a summary.

We analyze the time complexity of our approach by considering the two major computational steps: EWKM method (step 5) and the modularity maximization based method (step 9). The EWKM algorithm converges in a finite number of iterations [29]. The total computational complexity is $O(hmnk)$, where h, m, n, k denote the number of iterations, the number of social objects, dimensions of social objects, and the number of clusters respectively. The computational complexity increase almost linearly as h, m, n, k increase. As to the modularity maximization based method, the time complexity is $O(N_u log^2 N_u)$, where N_u represents the number of users/members involved in the social activities. Therefore, the time complexity of our approach is $O(hmnk + N_u log^2 N_u)$.

Why adopt these three main steps: social objects clustering; topic-based user partitioning, and link-based community detection? An alternative might be to do the link based community detection first and then partition users based on topics. Here we consider the latter as a candidate approach. We will compare the candidate approach with our approach. Two advantages of our approach can be observed: (1) keeping the completeness of connections within the same topic; (2) increasing the efficiency of link analysis.

Taking Fig. 4 as an example, we assume that the interested topics of each user have been known and labeled on each node. That is, user A and B are interested in topic $T1$; user E and F are interested in $T2$; user G and H are interested in $T3$; user C, D and I are interested in three topics including $T1$, $T2$ and $T3$.

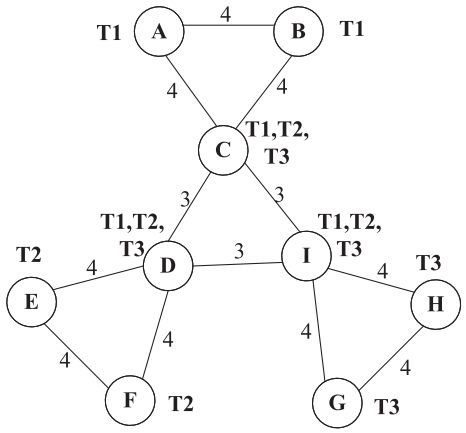


Fig. 4. An example of the social network. The node represents the user. The edge represents the social relations between users. The weight attached to each edge represents the strength of the communications. Labels attached to each user represents the corresponding user's interested topics.

Fig. 5 illustrates the processes and results of the candidate approach and our approach respectively. Different colors are used to represent different resulting communities (topical communities). From Fig. 5(a), we can see that the candidate approach generates nine topical communities on three topics: $T1 = \{\{A,B,C\},\{D\},\{I\}\}$, $T2 = \{\{D,E,F\},\{C\},\{I\}\}, T3 = \{\{G,H,I\},\{C\},\{D\}\}$. Our approach, however, detects three communities shown in Fig. 5(b): $T1 = \{A,B,C,D,I\}$, $T2 = \{C,D,E,F,I\}$, $T3 = \{C,D,G,H,I\}$. According to the results, we can conclude that our approach can keep the completeness of connections within the same topic.

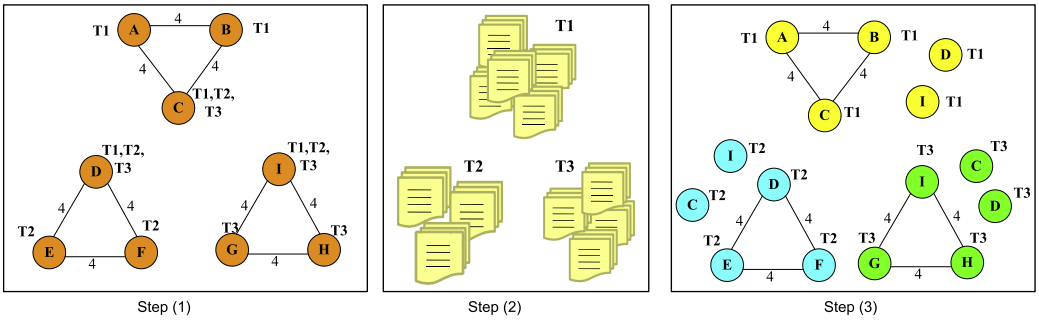
Another advantage of our approach is that it can increase the efficiency of link analysis. Performing the social objects clustering first can generate many sub-graphs (topical clusters) in step (3). Compared to the original user graph, the sub-graphs are often smaller, which can reduce the complexity of link analysis. Moreover, the link analysis can be done on these sub-graphs simultaneously, since each sub-graph represents a distinct topic.

Algorithm 3. Our approach-community detection based on social objects and link analysis

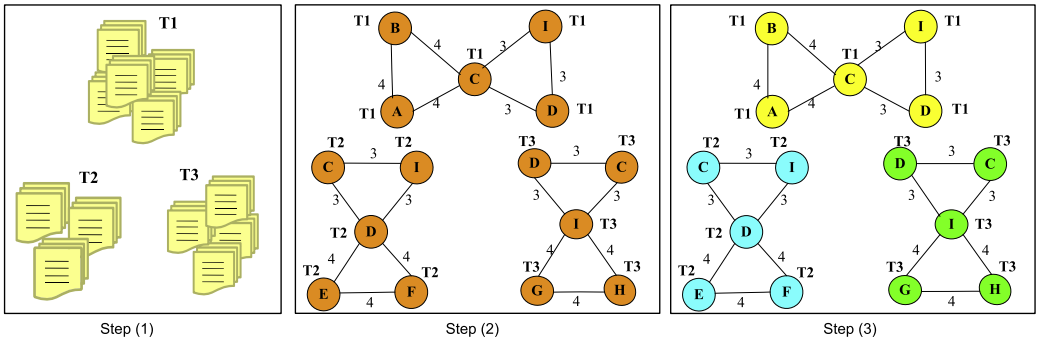
Input: social network data set,
the number of clusters for social objects K,
parameter γ,
the number of top dimensions for each cluster L;
Output: communities identified
1: construct the EG model for all the input data set;
2: preprocess the text objects;
3: represent each object with VSM and get the matrix denoting text objects set;
4: call EWKM algorithm (Algorithm 1) to cluster text objects;
5: select top L dimensions to label each cluster;
6: partition the members into different topical clusters according to the association relations between members and objects described in EG;
7: **for** (each topical cluster) **do**
8: call modularity maximization method (Algorithm 2);
9: **end for**
10: **return** topical communities detected;

4. Experiment and analysis

In this section, we present extensive experiments on real datasets to evaluate the performance of our approach. We first applied our approach to three datasets and detected topical communities.



(a) The whole process and the topical communities detected by the candidate approach. Step(1): link based partitioning; Step(2): social objects clustering; Step(3): topic based community detection;



(b) The whole process and the topical communities detected by the proposed approach. Step(1): social objects clustering; Step(2): topic based user partitioning; Step(3): link based community detection; detection;

Fig. 5. Explanations to why we use the above three steps with a simple example. The input social network is described in Fig. 4.

Then we compared the performance of our approach with that of the traditional modularity method. Before going into details, we first describe the datasets and introduce the performance metrics to be used in the experiments.

4.1. Experimental datasets

Three real datasets used in our experiments are described in the following:

Enron dataset: The enron dataset is an email corpus provided by the CALO Project (http://www.cs.cmu.edu/~enron/). This data set contains 275,332 emails from 151 users. To analyze the email network, we removed the pure senders and pure receivers from the email address list. Empty emails were also dropped. After this processing, we had 14,800 distinct emails belonging to 140 users. According to the topics given by Marti Hearst and her students (regulations, internal projects, company image, political influence, california energy crisis, internal company policy and operations, alliances and partnership, legal advice, talking points, meetings, trip reports) (http://bailando.sims.berkeley.edu/enron/enron_categories.txt), we clustered all the emails into 11 categories. That is, the number of topics was set to be $K = 11$ in our EWKM step.

Political Blogs Dataset: The political Blogs Dataset presents a blog network about the US political issues. The dataset was recorded in 2005 by Adamic and Glance [32]. There are totally 1,490 webblogs and 19,090 hyperlinks between them. Each blog is classified as conservative or liberal, represented as 0 for liberal, 1 for conservative. Therefore, the number of topics was set to be $K = 2$ in our EWKM step.

Cora dataset: The Cora dataset used here is a subset of the larger Cora citation dataset. There are totally 2708 nodes and 5429 links. Each node corresponds to one paper, while the link refers to the citation relation. All the publications included in the dataset are from 7 sub-categories of machine learning research: Case-based Reasoning, Genetic Algorithms, Neural Networks, Probabilistic Methods, Reinforcement Learning, Rule Learning and Theory. Therefore, the number of topics was set to be $K = 7$ in the EWKM step.

4.2. Performance metrics

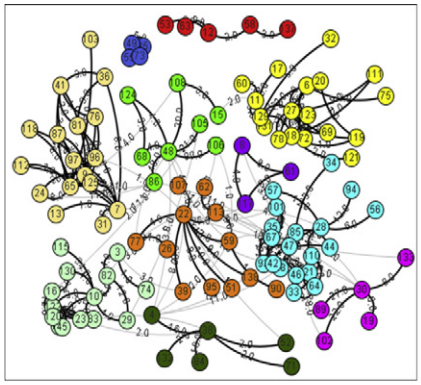
Considering that our approach combines the topic and link to detect communities, a reasonable evaluation metric should contain two aspects, one for topic and the other for linkage structure. That is, the expected results should keep each community's members with the same topic and strong connections.

Inspired by F-score, a well-accepted metric which considers both *precision* and *recall* to test the results in information retrieval, we define our performance evaluation metric as follows:

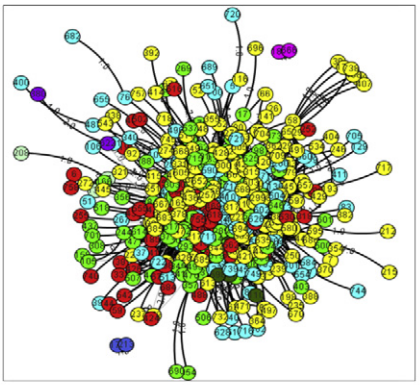
$$PurQ_\beta = (1+\beta^2)\cdot(Purity\cdot Q)/(\beta^2\cdot Purity + Q), \tag{6}$$

where

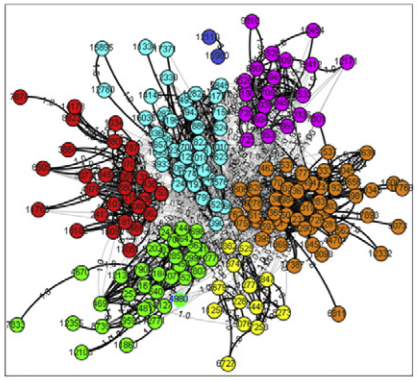
- *Purity* denotes the purity of topics in the detected communities and $Purity = 1/N_{cm}\cdot\sum_{i=1}^{i=N_{cm}}\max_{1\leqslant j\leqslant K}\{n_{ij}/n_i\}$, where N_{cm} denotes the number of resulting communities, n_{ij} refers to the number of nodes belonging to topic j and community i, n_i refers to the number of members in community i. The higher the *Purity*, the better the communities are partitioned from the perspective of topics.



(a) The topical communities detected in the cluster of internal company policies (a subtopic of Enron Email dataset).



(b) The topical communities detected in the cluster of liberal blogs.



(c) The topical communities detected in the cluster of reinforcement learning(a sub-category of Cora dataset).

Fig. 6. Samples of topical communities detected from the corresponding cluster of three different datasets. Different colors are used to represent different topical communities.

- Q denotes the modularity. It is often used to evaluate whether the division is good, in the sense that there are many edges within communities and only a few between them. That is, Q measures the communities from the perspective of the link structure. $Q = 1/K \cdot \sum_{p=1}^{p=K} Q^p$, where K is the number of topics in EWKM, (For the method that does not use EWKM, we assume $K = 1$). Q^p is the value of the modularity function for the topical cluster $Cluster^p$. The larger the Q, the better the communities are divided from the topological structure.
- β is a parameter to adjust the weight of *Purity* and Q. $\beta \in [0, \infty]$. $\beta = 1$ can be interpreted as that the metric is the harmonic mean of the *Purity* and Q. In that case, the purity of topics and the topology of the social network are considered being equal important. $1 < \beta < \infty$ means that compared to *Purity* the metric pays more attention to Q. For $0 < \beta < 1$, the metric puts more emphasis on *Purity* than Q.

According to the above definition, we can interpret $PurQ_\beta$ as a weighted value of *Purity* and Q. Generally speaking, the *Purity* increases with Q decreasing, and vice versa. The *Purity* will get its highest value if we divide the whole graph into n topics in which each node represents a unique topic. But in this situation, the Q value will become the lowest as no links are kept. On the contrary, the Q will achieve a higher value if we analyze the linkage for the whole graph. But in this situation, the *Purity* will be very low as each resulting community usually contains more than two topics. Therefore, our performance metric can make a balance between the *purity*

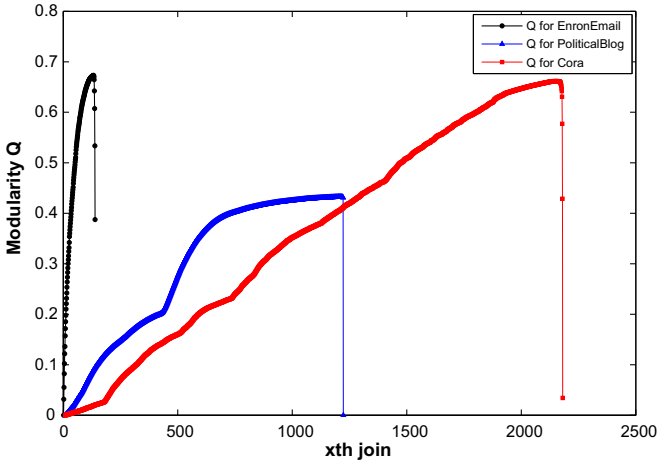
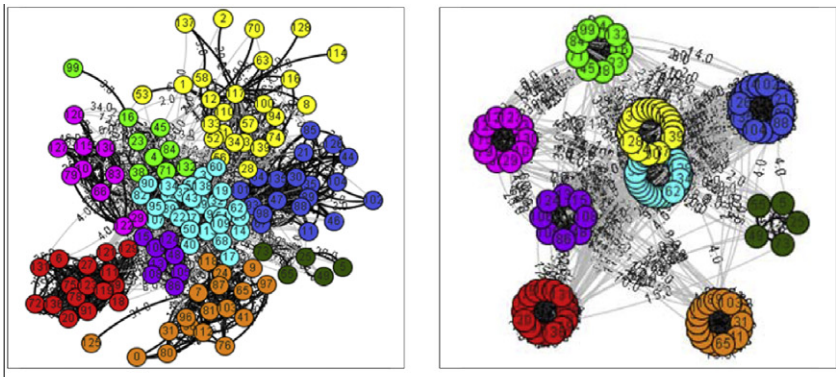
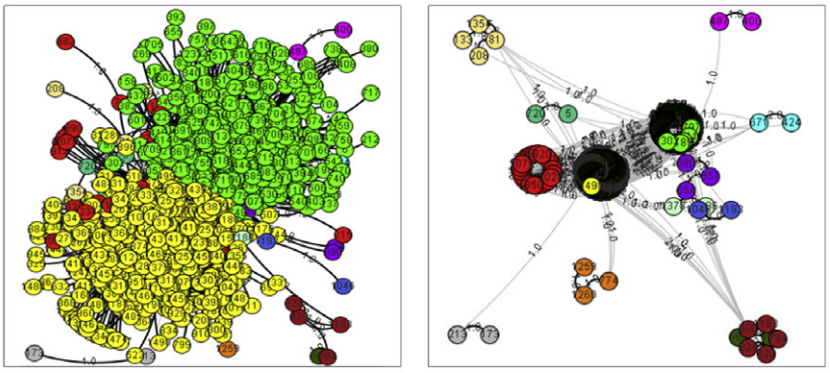


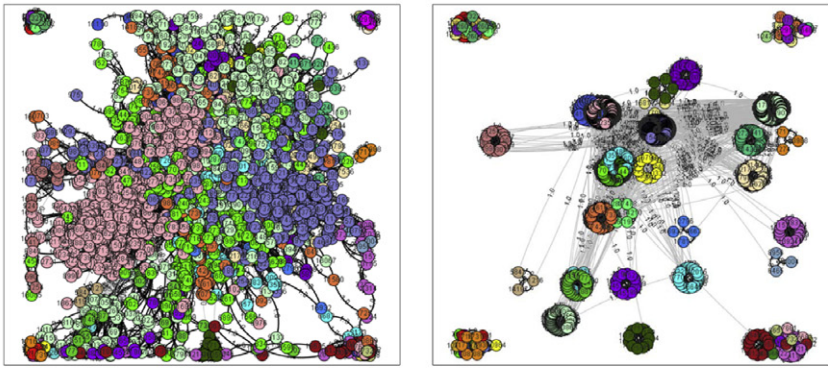
Fig. 7. Q value changes with the merge of communities for three networks: Enron Email, Political blogs, Cora. The x axis represents the number of merges and the y axis represents the corresponding Q values. The maximum Q for above three networks are 0.683114, 0.432282, 0.661694 respectively. The corresponding joins are: $x = 131$, $x = 1210$, $x = 2155$.



(a) 9 communities detected by modularity method within Enron Email dataset.



(b) 14 communities detected by modularity method within Political Blogs dataset.



(c) 62 communities detected by modularity method within Cora dataset.

Fig. 8. Communities identified by modularity maximization based method within 3 whole datasets: Enron Email dataset, Political Blog dataset, and Cora dataset. Different colors are used to represent different communities except sub-figure (c), since we designed only 16 kinds of colors in our program. For each sub-figure (a), (b) and (c), the right one is the circled layout of the left one, which can help us find the number of communities more easily.

and the Q to reduce the biased results. β here is used to adjust the emphasis of the above two aspects.

4.3. Experiment process and results

To analyze the above three datasets, the first step was to preprocess the datasets. As to the Enron Email dataset and Cora dataset, we considered the email-contents and paper-titles as the social objects. We adopted the vector space model to represent each mail or title. We first preprocessed the two datasets through splitting words, removing stop words and stemming each word to its rooting form. Then we computed the *tf*|*idf* value for each word. The words with very high or very lower *idf* values were removed.

Finishing the preprocessing steps above, we got two matrixes $M^{Mail}_{8514*4157}$ and $M^{Cora}_{2447*2334}$ which represent the Enron Email dataset and the Cora dataset. The rows of the matrix represent social objects while the columns represent the dimensions of unique words. As to the Political Blog Dataset, we took the blog contents as the social objects. Each blog had two kinds of values: liberal or conservative. Therefore, we represent the Political Blogs Dataset as matrix $M^{Blog}_{14790*2}$.

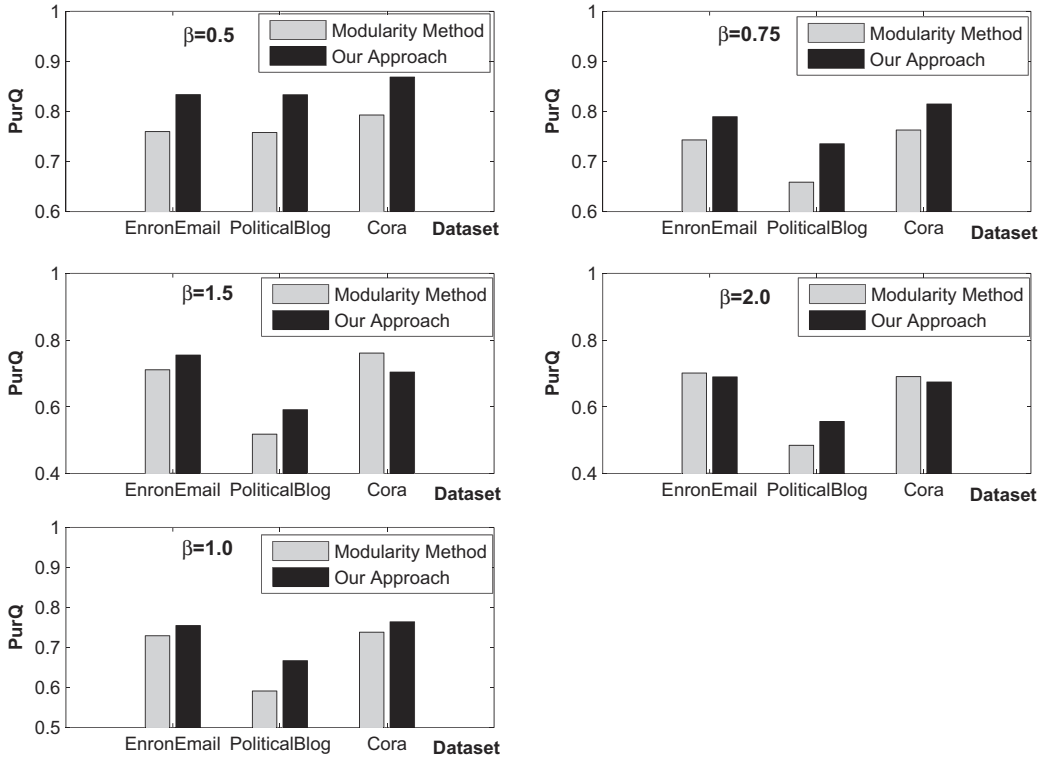


Fig. 9. The performance comparison under different parameter settings.

We performed the social objects clustering with the EWKM algorithm to find topics for each dataset with each matrix as input. Then we partitioned the members into different topical clusters and detected topical communities for each topical cluster. Some results are shown in Fig. 6.

With the Enron Email dataset, the cluster we selected is about internal company policies or operations (a subtopic of Enron Email dataset). The top 5 dimensions identified by EWKM are: confidenti (confidential), sampl (sample), bullet (bulletin), secreter (secretary), messag (message). We detected 11 topical communities in this cluster shown in Fig. 6(a). The average size of communities was 10.0909. The minimum size was 3 while the maximum size was 19.

For the Political Blogs Dataset, Fig. 6(b) shows the topical communities detected in the cluster of liberal blogs. Due to the speciality of the political blog dataset, there is only one dimension to describe its topic: liberal. We detected 9 topical communities with the average size of 63.6667. The minimum size was 2 while the maximum size was 201. Within these topical communities, there were two isolated topical communities {666,182} and {213,173}, although they have the same topic with others.

Fig. 6(c) shows the topical communities detected in the cluster of reinforcement learning (a sub category of Cora dataset). The top 5 dimensions identified by EWKM are: optimiz (optimization), learn (learning), heur (heuristic), adap (adaptive), rul (rules). We detected 7 topical communities in this cluster and the average size of communities was 29. The minimum size was 2 while the maximum size was 58. Within these topical communities, a special one was {12110,16980} which was isolated from others although they had the same topic.

4.4. Comparison and evaluation

In this subsection, we compare our results with the communities discovered by the modularity maximization based method. We use the metric defined in subSection 4.2 to evaluate the experimental results.

We first applied the modularity maximization based method to the Enron Email dataset, Political Blog dataset and Cora dataset. The Q value for each join or merge of communities in each dataset is illustrated in Fig. 7. For the Enron Email dataset, the Q value reaches its maximum at the 131st join, which corresponds to the 9 communities shown in Fig. 8(a). For the Political Blog dataset, the Q value reaches its maximum at the 1210th join, which results in 14 communities with the average size of 87.4286 shown in Fig. 8(b). For the Cora dataset, the Q value reaches its maximum at the 2155th join, which results in 62 communities with the average size of 35.7581 shown in Fig. 8(c).

We then used the $PurQ_\beta$ metric to evaluate the performance. In the experimental evaluation, β was set to be 0.5, 0.75, 1.0, 1.5, 2.0 respectively, which represent the different strengths for the topic and link. The corresponding results are shown in Fig. 9. We can see that our approach achieved a higher $PurQ$ than the modularity method in the cases of $\beta = 0.5$ and $\beta = 0.75$. In the case of $\beta = 1.5$, our approach got a low performance for the Cora dataset. But for other two datasets, our approach gained a high performance. When β increased to 2.0, our approach got a low or nearly equal performance compared to the modularity method. In the case of $\beta = 1.0$ which means the topic and link are considered to have the same weights, our approach achieved high performance for all three datasets. According to the explanation of β in Section 4.2, we can conclude that our approach have a better performance than the pure modularity based method, when the topic is at least as important as the link.

5. Conclusion

In this paper, we have proposed a topic oriented community detection approach based on social objects clustering and link analysis. Taking social objects into consideration, we first perform the object clustering with the Entropy Weighting K-Means algorithm. Then all the members involved in those social objects are partitioned into topical clusters, each of which represents a certain

topic. In order to differentiate the strength of connections, we perform a link analysis on each topical cluster and detect the topical communities. The modularity function is also employed to determine the appropriate number of communities. To evaluate the performance, we conducted experiments on three real data sets. Experimental results have shown that our approach gained a better performance than the traditional modularity based method, when the topic was at least as important as the link. Furthermore, the topical communities detected by our approach were more meaningful since they were empowered by topics.

Our approach has many potential applications. It can be applied to many kinds of social networks, which contain social objects. With the communities detected by our approach, we are able to improve the efficiency of collaborative learning, direct marketing, expert finding, and knowledge sharing for each topic, and make full use of collective intelligence.

Acknowledgement

This research is partly supported by Knowledge Innovation Project of Chinese Academy of Sciences under Grant No. KGCX2-YW-131, Shenzhen New Industry Development Fund under Grant No. CXB201005250021A. We thank Chao Li and Shuang Wang for their critical reading and careful revisions of this manuscript. We greatly appreciate the reviewers and editor for their valuable suggestions and comments to improve this work.

References

[1] J. Leskovec, K. Lang, M. Mahoney, Empirical Comparison of Algorithms for Network Community Detection, in: Proceedings of the 19th International Conference on World Wide Web (WWW), 2010, pp. 631–640.
[2] S. Fortunato, Community detection in graphs, Physics Reports 486 (2010) 75–174.
[3] Z. Xia, Z. Bu, Community detection based on a semantic network, Knowledge-Based Systems 26 (2012) 30–39.
[4] W. Yuan, D. Guan, Y.-K. Lee, S. Lee, S.J. Hur, Improved trust-aware recommender system using small-worldness of trust networks, Knowledge-Based Systems 23 (3) (2010) 232–238.
[5] F. Wu, B. Huberman, L. Adamic, J. Tyler, Information flow in social groups, Physica A: Statistical Mechanics and its Applications 337 (1-2) (2004) 327–335.
[6] P. Liu, B. Raahemi, M. Benyoucef, Knowledge sharing in dynamic virtual enterprises: a socio-technological perspective, Knowledge-Based Systems 24 (3) (2010) 427–443.
[7] S. Smyth, A spectral clustering approach to finding communities in graphs, in: Proceedings of the 5th SIAM International Conference on Data Mining, 2005, pp. 76–84.
[8] R. Guimera, L. Amaral, Functional cartography of complex metabolic networks, Nature 433 (7028) (2005) 895–900.
[9] G. Flake, S. Lawrence, C. Giles, F. Coetzee, Self-organization of the web and identification of communities, Communities 35 (3) (2002) 66–71.
[10] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5) (1999) 604–632.
[11] M. Girvan, M. Newman, Community structure in social and biological networks, Proceedings of the National Academy of Sciences of the United States of America 99 (12) (2002) 7821–7826.
[12] F. Wu, B. Huberman, Finding communities in linear time: a physics approach, The European Physical Journal B-Condensed Matter and Complex Systems 38 (2) (2004) 331–338.
[13] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[14] B. Yang, W. Cheung, J. Liu, Community mining from signed social networks, IEEE Transactions on Knowledge and Data Engineering 19 (10) (2007) 1333–1348.
[15] J. Zeng, S. Zhang, C. Wu, A framework for WWW user activity analysis based on user interest, Knowledge-Based Systems 21 (8) (2008) 905–910.
[16] A. McCallum, A. Corrada-Emmanuel, X. Wang, Topic and role discovery in social networks, in: Proceedings of the 19th international joint conference on Artificial intelligence, 2005, pp. 786–791.
[17] Y. Tian, R. Hankins, J. Patel, Efficient aggregation for graph summarization, in: Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008, pp. 567–580.
[18] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, A. Barabási, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.
[19] R. Guimerà, M. Sales-Pardo, L. Amaral, Module identification in bipartite and directed networks, Physical Review E 76 (3) (2007) 036102.
[20] E. Leicht, M. Newman, Community structure in directed networks, Physical Review Letters 100 (11) (2008) 118703.
[21] S. Lehmann, M. Schwartz, L. Hansen, Biclique communities, Physical Review E 78 (1) (2008) 016108.
[22] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, Y. Fan, Clustering coefficient and community structure of bipartite networks, Physica A: Statistical Mechanics and its Applications 387 (27) (2008) 6869–6875.
[23] Y. Kim, S. Son, H. Jeong, Link Rank: Finding communities in directed networks, Physical Review E 81 (1) 016103.
[24] M. Newman, Modularity and community structure in networks, Proceedings of the National Academy of Sciences of the United States of America 103 (23) (2006) 8577–8582.
[25] M. Barber, Modularity and community detection in bipartite networks, Physical Review E 76 (6) (2007) 066102.
[26] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (ACM SIGIR), 1999, pp. 50–57.
[27] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research 3 (2003) 993–1022.
[28] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August, 2004, pp. 22–25.
[29] L. Jing, M. Ng, J. Huang, An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data, IEEE Transactions on Knowledge and Data Engineering 19 (8) (2007) 1026–1041.
[30] A. Clauset, M. Newman, C. Moore, Finding community structure in very large networks, Physical Review E 70 (6) (2004) 066111.
[31] M. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69 (2) (2004) 026113.
[32] L. Adamic, N. Glance, The political blogosphere and the 2004 US election: divided they blog, in: Proceedings of the 3rd International Workshop on Link Discovery, 2005, pp. 36–43.