Journal of King Saud University – Computer and Information Sciences (2016) **28**, 303–310

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Topic-oriented community detection of rating-based social networks

Ali Reihanian [a,*], **Behrouz Minaei-Bidgoli** [b], **Hosein Alizadeh** [b]

[a] *Department of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran*
[b] *Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran*

Received 30 April 2015; revised 12 June 2015; accepted 3 July 2015
Available online 2 November 2015

KEYWORDS

Content analysis;
Topical community;
Community detection;
Modularity;
Purity

Abstract Nowadays, real world social networks contain a vast range of information including shared objects, comments, following information, etc. Finding meaningful communities in this kind of networks is an interesting research area and has attracted the attention of many researchers. The community structure of complex networks reveals both their organization and hidden relations among their constituents. Most of the researches in the field of community detection mainly focus on the topological structure of the network without performing any content analysis. In recent years, a number of researches have proposed approaches which consider both the contents that are interchanged in networks, and the topological structures of the networks in order to find more meaningful communities. In this research, the effect of topic analysis in finding more meaningful communities in social networking sites in which the users express their feelings toward different objects (like movies) by means of rating is demonstrated by performing extensive experiments.

1. Introduction

With the advance of information technology, online communications between people have increased significantly. This kind of communications have become more organized subsequent to the emergence of social networks. For example, folksonomies are social tagging sites which their users collaboratively express their feelings and sentiments toward a special resource like a movie or music by means of descriptive keywords (tags) (Chakraborty et al., 2012) or ratings. Finding meaningful communities in this kind of networks is an interesting research area and has attracted the attention of many researchers. The community structure of complex networks reveals both their organization and hidden relations among their constituents (Lancichinetti and Fortunato, 2012). A community (also sometimes referred to as a module or cluster (Leskovec et al., 2010)) is a dense sub network within a larger network, such as a close-knit group of friends in a social network or a group of interlinked web pages on the World Wide Web (Newman, 2011). As the people in the same community may usually have common hobbies and social functions, the identified communities can be used in

* Corresponding author.
E-mail address: areihanian@ustmb.ac.ir (A. Reihanian).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2015.07.001

collaborative recommendation, information spreading, knowledge sharing and other applications that are beneficent to us (Zhao et al., 2012).

Most of the researches in the field of community detection mainly focus on the topological structure of a network. They just build a network of individuals without performing any content analysis. Most of these networks are built based on the number of communications between individuals. Actually, these researches just consider the graph structure of a network for finding communities and no content analysis has been used in the process of their proposed approaches.

Despite of the original definition of the networks, nowadays, real world networks contain a vast range of information including shared objects, comments, following information, etc. It is unreasonable for a community to be explained by a single entity because the community members are generally interacting with each other via a large number of distinguishable ways in various domains.

One of the possible solutions is to find topical clusters in which the nodes have the same topic of interest. Each topical cluster represents one of the topics of interest in the network. Then, a community detection algorithm can be applied to these topical clusters to find the ultimate communities (Zhao et al., 2012). In this way, we can analyze and estimate the effect of topic consideration in community detection.

In this paper, the effect of topic analysis in finding more meaningful communities in social networking sites in which the users express their feelings toward different objects (like movies) by means of rating, is demonstrated by performing extensive experiments. Therefore, the network is partitioned into different topical clusters in which the nodes have the same topic of interest. Then, a community detection algorithm is applied to the topical clusters in order to find more meaningful communities. This will lead us to communities in which the nodes are tightly connected and have the same topic of interest. This process is called topic-oriented community detection (Zhao et al., 2012). At last, the results of community detection with topic consideration are compared with the results of community detection without considering the topics of interest. Quantitative evaluations reveal that the results of community detection will be improved when the topic of interest in the network is considered.

The remainder of the paper is outlined as follows. Section 2 explains the motivation of our research. In Section 3, related works are reviewed. Section 4 explains the topic-oriented community detection. In order to evaluate the effect of topic consideration in identifying communities of rating-based social networks, extensive experiments are conducted on real-life data sets. The descriptions of these data sets, the experimental results and their analyses are given in Section 5. Finally, the conclusions are given in Section 6.

2. Motivation

In this section, the motivation of our research is explained with an example. Look at the example illustrated in Fig. 1. Fig. 1(a) is a network of 8 nodes and 11 edges. We call this network a basic network. Each node is an individual in the network, and each edge is the social relation of interactions or communications. The weight of each edge represents the number of communications between the related nodes. For example, if node i finishes five communications with node j, the assigned weight of their related edge will be 5. Consider that the topics of interest for each node are assigned to them manually. These topics represent the domain of interest for each individual in the network. In this specific network, each node can be interested in discussions related to religion, irreligion or both of them.

Fig. 1(b) shows the identified communities after applying a community detection algorithm on the basic network. In this situation, no content analysis has been performed. The members of each identified communities are connected, but as you can see, the community that is located at the top of the Fig. 1(b) incorporates different topics. Two members in this community are interested in religion while three members are interested in irreligion.

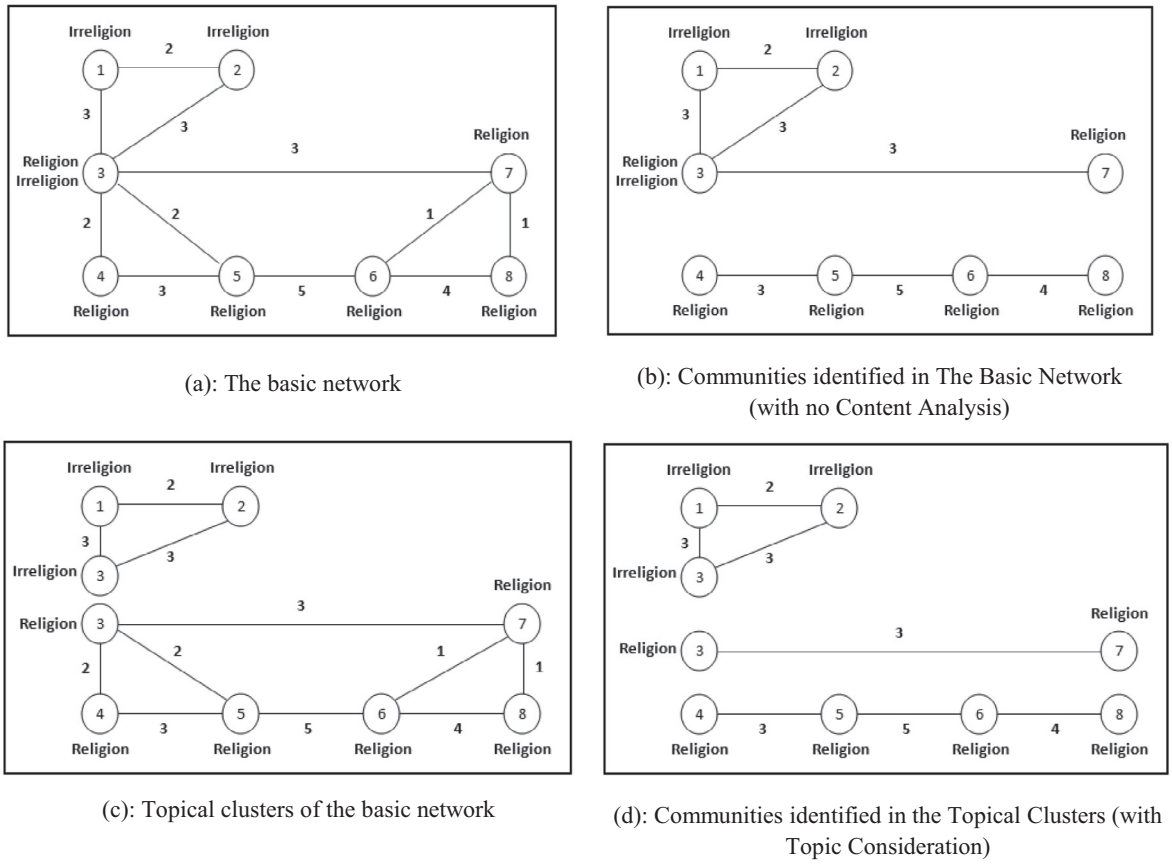
Zhao et al. (2012) extracted topical clusters from the basic network in order to detect communities which have a unique topic of interest and connected members. Each topical cluster contains the nodes of the basic network which have the same topic of interest. Fig. 1(c) shows the partition of the network that has two topical clusters. For example, in the topical cluster that is located at the bottom of the figure, all of the members are interested in the discussions related to religion. Then, a community detection algorithm will be applied to each topical cluster. Fig. 1(d) shows the identified communities in this situation. Each community has members who are connected to each other and have the same topic of interest. This is the condition we aim to analyze in the rating-based social networks.

3. Related works

Many researches have been done in the area of community detection. Most of these researches mainly focus on the topological structure or linkage patterns of networks. They merely consider the graph structure of a network for finding communities, while no content analysis is used in the process of their proposed approaches.

According to the community detection strategies which were employed in these researches, their proposed methods can be classified into optimization-based methods and heuristic methods. Some of the optimization-based methods focus on optimizing an objective function (Zhao et al., 2012). One of the most important works in the literature was a research done by Newman and Girvan, in which they introduced modularity as an objective function (Newman and Girvan, 2004). A large amount of works have been done to optimize modularity such as the methods which were developed by Arenas et al. (2007), Leicht and Newman (2008), Newman (2004). This function has been influential in the literature of community detection, and has gained success in many applications. Modularity is used to evaluate the quality of a particular division of a network into communities (Zhao et al., 2012). On the other hand, heuristic methods such as the GN algorithm (Girvan and Newman, 2002) and the CPM algorithm (Palla et al., 2005) design a graph clustering algorithm based on intuitive assumptions (Zhao et al., 2012).

Even though these researches have gained success in some applications, since they mainly focus on the topological structure of the networks, they ignored the contents interchanged between members. As a result, the relationships between the



(a): The basic network

(b): Communities identified in The Basic Network (with no Content Analysis)

(c): Topical clusters of the basic network

(d): Communities identified in the Topical Clusters (with Topic Consideration)

Figure 1 An example illustrating the motivation of our research.

members in these researches are mainly based on the total number of communications.

Another group of researches tend to partition the networks into different groups of nodes in which every node has the same topic of interest. In other words, these researches focus on topic modeling through analyzing the contents of social objects. It should be considered that social objects refer to the objects, like e-mails, which people communicate with each other through them. Several topic models have been proposed such as LSA (Deerwester et al., 1990), pLSA (Hofmann, 1999), LDA (Blei et al., 2003), etc. Latent semantic analysis (LSA) is a widely adopted approach to map the high dimensional co-occurrence matrix into a lower dimensional representation as latent semantic space to reveal semantic relations between entities. Hofmann (1999) made the significant leap forward to LSA and proposed the probabilistic LSA (pLSA) where the detected clusters are more topic-oriented. Blei et al. (2003) proposed the Latent Dirichlet Allocation (LDA), a three-level hierarchical Bayesian model that models words and documents over an underlying set of topics, to avoid the pLSA's serious problems of over-fitting (Ding, 2011).

As mentioned earlier, these researches' goal is to find communities in which all members have the same topic of interest, while they ignore the relationships between members. As a result, communities detected by these researches tend to contain topologically-diverse sub-communities within each community (Ding, 2011).

In recent years, a number of researches have proposed approaches which consider both the contents that are interchanged in the networks and the topological structures of the networks in order to find more meaningful communities. Zhao et al. (2012) proposed a topic-oriented community detection approach based on social objects' clustering and link analysis. Their proposed approach could identify the topical communities which reflect the topics and strengths of connections simultaneously. Zhu et al. (2013) combined classic ideas in topic modeling with a variant of mixed-membership block model which is recently developed in the statistical physics community. In their research, Zhu et al. combined topic-modeling with link structure. Zhao and Ma (2012) proposed a framework to apply a semantically structured approach to the Web service community modeling and discovery.

4. Topic-oriented community detection in a social network

As we mentioned earlier, the goal of this paper is to demonstrate the effect of topic consideration in finding more meaningful communities in social networking sites in which the users express their feelings toward different objects (like movies) by means of rating. For this cause, some components of the frame work which was proposed by Zhao et al. (2012) are changed in order to be applicable to the mentioned social networks. This framework detects communities which have a unique topic of interest and connected members. Each

community contains the nodes of the network which have the same topic of interest. This framework is implemented in four steps: preprocessing and annotating topic labels, clustering social objects, creating topical clusters, and applying a community detection algorithm to the topical clusters.

4.1. Preprocessing and annotating topic labels

In this step, data sets are preprocessed and ready to use. In this process, the social objects are recognized. Generally, people communicate with each other through social objects. These objects often imply the topics which people are interested in. Social objects can be classified into two kinds of situations (Zhao et al., 2012): (1) the social objects which are attached to multi-members, (2) the social objects which are attached to one member.

In the first situation, the edges between members are built because of a social object. An example of this situation can be happened in a movie rating network. In this network, edges between members are built when they rate the same movie. As a matter of fact, in this network, each movie (social object) is attached to multi members. The members of the movie rating network are connected to each other due to the rating of the same movie.

In the second situation, each social object is attached to only one member. Therefore the social objects are considered to be the attributes of members of the network. An example of this situation can be found in a paper citation network. In this network, papers (members) cite each other. Also, each paper contains a text content (the title of a paper) which is a social object and can be considered as the attribute of the corresponding paper.

Fig. 2 shows the two different kinds of relations between the members of a network and social objects. The network which is located in the left side of the Fig. 2 is a movie rating network. As it is clear, the edges between members are built because of the social objects. Also, the network which is located on the right side of the Fig. 2 is a paper citation network. In this network, each social object is the attribute of its corresponding paper. Since in this paper the social networking sites in which the users express their feelings toward different objects are analyzed, the first situation is encountered.

So, in this step, data sets are preprocessed and ready to use. In this process, the social objects are recognized. Afterward, the topics of each social object in a data set are retrieved. Subsequently, each social object is labeled by its corresponding topic. In some cases the topics of each social object can easily be retrieved manually, or there are corresponding tags which represent the topics for each social object. But in cases where a social object is represented by text and its labels cannot easily be retrieved, a method has been introduced by Zhao et al. (2012) which can annotate the topic label to each social object.

4.2. Clustering social objects

In this step, social objects in a network are partitioned into different clusters. Each cluster represents a unique topic which is shared by its members. In other words, according to their labeled topics, social objects are partitioned into different clusters in a way that each cluster includes members with the same topic. Since the data sets which are used in this paper contain social objects with labeled topics, we manually partition these social objects into different clusters.

4.3. Creating topical clusters

Using the results that are generated in the previous step, we partition the members of the network into different topical clusters. In the first step, each social object has been annotated with a topic label. In this step, members are partitioned into different topical clusters considering the topic labels of the social objects they are involved in. Thus in this step we find clusters in which every member has the same topic of interest. Therefore the total number of topical clusters is equal to the number of topics of interest in the network. A user can be a member of several topical clusters, since it is common for a user to be interested in several topics.

4.4. Applying a community detection algorithm to the topical clusters

This step aims to find communities in each of the topical clusters which were created in the previous step. Members in each topical cluster are connected to each other with different strengths. Based on the number of ratings on the same social objects, some members may have stronger connections, while some others may have weak or no connections. This has been concluded according to the topic analysis that has been performed in the framework. Since the result of the framework is to detect communities which have a unique topic of interest and connected members, we should apply a community detection algorithm to the previously created topical clusters in order to identify the tightly connected members.

In order to perform this process, many community detection algorithms can be employed such as GN and so on. Newman proposes an important algorithm to partition network graphs of links and nodes into sub graphs. He also introduces a concept which is called modularity. In the case of weighted networks, modularity has been defined as follows (Newman, 2004):

$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{1}$$

where A_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i, c_i is the community to which the vertex i is assigned, the δ function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and also $m = \frac{1}{2}\sum_{ij} A_{ij}$.

Since Newman's algorithm was very time-consuming, Blondel et al. (2008) suggest the modified version of the algorithm in order to make it faster, giving rise to what is known as the "Louvain method". This algorithm is a modularity maximization algorithm which iteratively optimizes the modularity in a local way and aggregates nodes of the same community (Wang et al., 2014). In this paper, the "Louvain method" has been applied in order to find topical communities.

4.5. Application of the topic-oriented community detection framework to different kinds of social networks

As it was mentioned earlier, the data sets which were used in this paper are related to the social networking sites in which

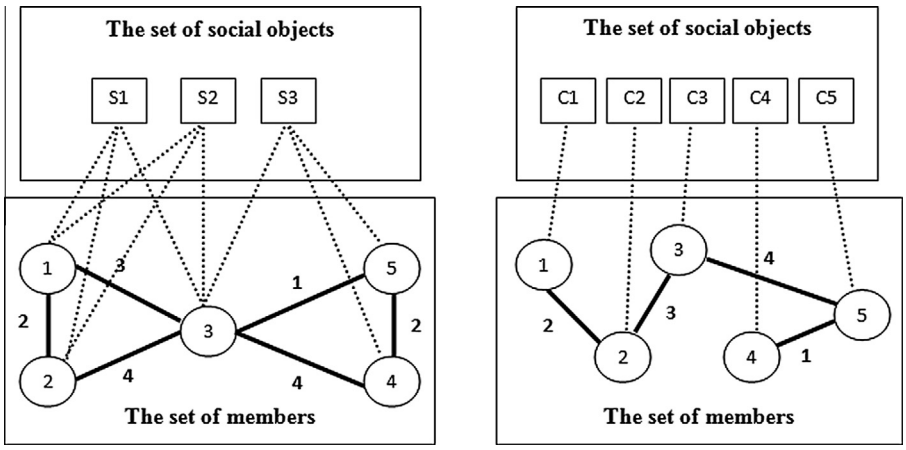


Figure 2 Two different kinds of relations between the members of networks and social objects.

the users express their feelings toward different objects by means of rating. Thus, the topic-oriented framework in its current format, which is explained in this section, can be applied to this kind of social networks. But, consider that we want to apply the topic-oriented framework to a typical email network or other kinds of social networks in which no topics of interest and ratings are available. In order to apply the topic-oriented framework to these kinds of social networks, one step of the framework should be applicable to those networks: clustering social objects.

Different methods can be used to perform the social object clustering according to the type of social objects. For example, a novel method has been proposed by Zhao et al. (2012) to cluster the text social objects like emails. This method combines the vector space model with the Entropy Weighting K-Means (EWKM) Jing et al. (2007) in order to cluster the text social objects. On the other hand, sentiment analysis of communications' content can be performed in order to compute the weight of each semantic relationship. The mentioned methods are applicable to the contents in form of texts. But nowadays, social networks contain many other contents with different natures, like images, sounds and etc, which are interchanged between individuals. In our future works, we have a plan to introduce a framework which can detect meaningful communities in these kinds of networks.

5. Experiment and analysis

In this section, the results of our research are presented. First, five real-life data sets along with a performance metric which were used in the experiments are described. Then, the process of detecting topical communities in the mentioned data sets is discussed and its results are analyzed. Finally, the results of topic-oriented community detection (with performing content analysis) are compared with the results of community detection without performing any content analysis.

5.1. Real life data sets and performance metric

We used the publicly available data sets in our experiments which are Movielens 100k, Book-Crossing, CIAO, MovieTweetings and Movielens Latest. Movielens 100k is a rating data set which is collected from the Movielens web site (http://movielens.org). This data set is published by Grouplens research group (http://grouplens.org) and consists of 100,000 (100k) ratings from 943 users which were given to 1682 movies. Book-Crossing data set (Ziegler et al., 2005) is a rating data set which is collected from the Book-Crossing community (http://www.bookcrossing.com). It contains 278,858 users providing 1,149,780 ratings to 271,379 books. CIAO data set (Tang et al., 2013a,b; Tang et al., 2012a,b) is a rating data set which is collected from a product review site (http://ciao.com) in which users share their opinions about a product by means of rating or commenting. There are 35,773 ratings in this data set which are attached to 16,850 products by 2248 users. MovieTweetings (Dooms et al., 2013) is a data set consisting of ratings on movies that were contained in well-structured tweets on Twitter. In our experiment, we used the latest version of this data set which contains 389,735 ratings from 37,048 users given to 21,179 movies. The last data set which was used in our experiment is the latest version of Movielens data set which is called the Movielens Latest. This data set, which is collected in 2015, consists of 706 users providing 100,023 ratings to 8552 movies.

As described earlier in this paper, the topic-oriented community detection framework considers the results of topic analysis for finding more meaningful communities. So, in order to evaluate this framework, two aspects should be considered: topic and linkage structure. It means that the expected results should keep each community's members with the same topic and strong connections. Zhao et al. (2012) introduced a performance evaluation metric which considers both topic and linkage structure. This metric has been defined as follows:

$$\mathrm{Pur}Q_\beta = (1+\beta^2)(\mathrm{Purity}\cdot Q)/(\beta^2\cdot \mathrm{Purity}+Q) \tag{2}$$

As it is clear in the above equation, The $\mathrm{Pur}Q_\beta$ has three parameters which are Q, Purity and β. Q denotes the modularity. This parameter measures the communities from the perspective of link structure. The larger the Q, the better the communities are divided from the perspective of topological structure. In our experiment, for each topical cluster, modularity is calculated by Eq. (1). Since the topic-oriented framework may generate more than one topical cluster for each data set, the total value of modularity in this framework is calculated as follows:

$$Q = \sum_{i=1}^{n} \frac{\text{Weight}_{\text{TC}_i}}{\text{Weight}_T} \cdot Q_{\text{TC}_i} \tag{3}$$

where n is the number of generated topical clusters. Q_{TC_i} is the value of modularity for the topical cluster TC_i. $\text{Weight}_{\text{TC}_i}$ is the sum of the weights of edges in the topical cluster TC_i. Weight_T is the sum of the weights of edges in the topical cluster, which is directly created from the basic network (when no topical clustering has been performed). It should be considered that since in this framework no communications' content analysis is performed, the weight of each relationship between two members is the number of ratings which are given to the same social objects by these two members.In Eq. (2), Purity represents the purity of topics in the detected communities and is calculated as follows (Zhao et al., 2012):

$$\text{Purity} = 1/N_{\text{cm}} \cdot \sum_{i=1}^{N_{\text{cm}}} \max_{1 \leqslant j \leqslant k} \{n_{ij}/n_i\} \tag{4}$$

where N_{cm} represents the number of detected communities, n_{ij} refers to the number of nodes belonging to topic j, and community i, n_i refers to the number of nodes in community i. k is the number of topics in the network. The higher the Purity, the better the communities are partitioned from the perspective of topics.

β is a parameter to adjust the weight of Purity and Q and $\beta \in [0, \infty]$. If we consider the purity of topics and the topology of the network equally important, the value of β should be set to 1. If we want to pay more attention to Purity in comparison with Q, then the value of β should be set to a number between 1 and ∞. On the other hand, if we want to pay more attention to Q in comparison with Purity, the value of β should be set to a number between 0 and 1. Actually β is used in Eq. (2) to adjust the emphasis of topics and link structure (Zhao et al., 2012).

5.2. Experiments

In order to identify the communities by applying the topic-oriented community detection framework to the five introduced data sets, four steps (according to Section 4) have been taken. The first step was to preprocess the data sets. As to the Movielens 100k, Book-Crossing, MovieTweetings and Movielens Latest data sets, movies and books were considered as the social objects. So, for the Movielens 100k, MovieTweetings and Movielens Latest data sets, the genres of the movies were extracted. These extracted genres are the same as the genres attached to each movie by IMDB (http://www.imdb.com). Then, for the Movielens 100k data set, all the movies which were in the genres of Documentary or Western were retrieved. As you know, the genre of a movie represents the general topic in which a movie is made about. For the MovieTweetings data set, all the movies which were in the genres of Short or Documentary were retrieved. For the Movielens Latest data set, all the movies which were in the genres of Animation or Musical were retrieved. For the Book-Crossing data set, we extracted the categories of 93 books from Amazon (http://www.amazon.com). As for the CIAO data set, products were considered as the social objects. Each product's category was attached to it in the data set. Thus for the Book-Crossing data set and the CIAO data set, the categories represent the topics of each product or book.

The second step was to cluster the social objects. As for the Movielens 100k data set, the movies were partitioned into two clusters of Documentary and Western. The Documentary cluster contained 50 movies while the Western one contained 27 movies. As for the Book-Crossing data set, the books were partitioned into two clusters of Fiction and Non-Fiction. The Fiction cluster contained 80 books, while the Non-Fiction cluster contained 13. The products in the CIAO data set were partitioned into six clusters of DVDs, Books, Beauty, Music, Travel, and Food and Drink. The DVDs cluster contains 2057 products, The Books cluster contains 2803 products, the Beauty cluster contains 2333 products, the Music cluster contains 1801 products, the Travel cluster contains 3922 products and finally the Food and Drink cluster contains 3937 products. The movies in the MovieTweetings data set were partitioned into two clusters of Short and Documentary. The Short cluster contained 718 movies while the Documentary cluster contained 1334 movies. Finally for the Movielens Latest data set, the movies were partitioned into two clusters of Animation and Musical. The Animation cluster contained 339 movies while the Musical one contained 315 movies.

The third step was to create topical clusters. Therefore in each data set, the users who rate the social objects in each cluster were partitioned into topical clusters. For example, all users who rate the movies in the cluster of "Documentary" were partitioned into the topical cluster of "Documentary". The members of each topical cluster rated the social objects which have the same topics. Thus according to the number of topics, we achieved two topical clusters for the Movielens 100k, Book-Crossing, MovieTweetings and Movielens Latest data sets and 6 topical clusters for the CIAO data set. As mentioned earlier, since in this framework no communications' content analysis is performed, the weight of each relationship between two members is the number of ratings which are given to the same social objects (for example, two movies in the genre of Documentary) by these two members.

The last step was to detect topical communities. Thus we applied the "Louvain method" to each topical cluster created in the previous step. In order to accurately calculate the modularity, we applied the Louvain method to each topical cluster ten times, and calculated the average of the achieved values of modularity.

Table 1 gives the results achieved by applying the topic-oriented community detection framework to the Movielens 100k, Book-Crossing, CIAO, MovieTweetings and Movielens Latest data sets. In this table, the columns "Topical Clusters", "No. of Edges" and "No. of Nodes" represent the created topical clusters in the process of applying the topic-oriented framework to the five mentioned data sets, the number of edges and the number of nodes existing in each of these topical clusters, respectively. Moreover, the columns "Total Modularity" and "Purity" denote the overall modularity value (Q) and Purity value for all of the topical communities.

As it is clear in Table 1, Purity has its maximum value in each of the five data sets. The reason is that, the topical clusters created in each data set incorporate members which are interested in the same unique topics. Therefore the purity of topics in each of the topical clusters is 1 according to Eq. (4). It should be considered that it is possible for a certain user to be in several topical clusters, since the interest of people in several different topics is common. Thus some of the members of topical clusters in each data set may be the same. For example,

Table 1 The results achieved by applying the topic-oriented community detection framework to Movielens 100k, Book-Crossing, CIAO, MovieTweetings and Movielens Latest data sets.

Data sets	Topical clusters	No. of edges	No. of nodes	Total modularity	Purity
Movielens 100k	Documentary	15,833	352	0.1244	1
	Western	69,369	491		
Book-crossing	Fiction	8531	1021	0.8469	1
	Non-Fiction	1587	191		
CIAO	DVDs	53,916	1356	0.3086	1
	Books	8999	904		
	Beauty	5267	811		
	Music	2076	569		
	Travel	12,905	867		
	Food & Drink	29,763	1193		
MovieTweetings	Short	1667	352	0.5111	1
	Documentary	116,880	2640		
Movielens Latest	Animation	80,149	601	0.1732	1
	Musical	61,515	573		

Table 2 Comparison of modularities, Purities and PurQ_βs which were achieved by applying the topic-oriented framework along with the classical community detection framework to each of the five mentioned data sets.

Data set	Frameworks	Total modularity	Total purity	PurQ_β				
				$\beta = 0.5$	$\beta = 0.75$	$\beta = 1$	$\beta = 1.5$	$\beta = 2$
Movielens 100k	Classical	0.1086	0.9777	0.3760	0.2519	0.1955	0.1495	0.1321
	Topic-oriented	0.1244	1	0.4154	0.2830	0.2213	0.1703	0.1509
Book-Crossing	Classical	0.8375	0.9050	0.8906	0.8795	0.8699	0.8572	0.8502
	Topic-oriented	0.8469	1	0.9651	0.9389	0.9171	0.8888	0.8737
CIAO	Classical	0.2899	0.8279	0.6038	0.4963	0.4294	0.3624	0.3332
	Topic-oriented	0.3086	1	0.6906	0.5535	0.4716	0.3920	0.3581
MovieTweetings	Classical	0.5067	0.9912	0.8321	0.7374	0.6706	0.5964	0.5616
	Topic-oriented	0.5111	1	0.8394	0.7438	0.6765	0.6016	0.5665
Movielens Latest	Classical	0.122	0.9532	0.4034	0.2761	0.2163	0.1667	0.1478
	Topic-oriented	0.1732	1	0.5116	0.3678	0.2953	0.2323	0.2075

consider the case that a user rated several different movies. Some of these movies were in the genre of Documentary, and the others were in the genre of Western. Therefore this user belongs to both topical clusters in the Movielens 100k data set.

5.3. Comparison

In order to prove the superiority of the results of detecting communities with topic consideration, in this section, we compare the results of topic-oriented community detection, which was implemented in Section 5.2, with the results of classical community detection in which no content analysis is performed.

In the process of classical community detection approach, a community detection algorithm is applied to a network in which the weight of the edges represents the number of communications between relevant nodes. In this condition, no content analysis is done.

We first applied the "Louvain method" to the basic networks of the Movielens 100k, Book-Crossing, CIAO, MovieTweetings and Movielens Latest data sets (implementing the classical community detection framework). Then we partitioned the basic networks of the five mentioned data sets into topical clusters. Each topical cluster includes members which have the same topic. Afterward, the Louvain method was applied to these topical clusters (implementing the topic-oriented community detection framework which was discussed in Section 5.2). We then used PurQ_β to evaluate the performances in the experimental evaluation. The corresponding results are given in Table 2. Consequently, as it is shown in Table 2, β was set to 0.5, 0.75, 1, 1.5, 2 respectively, which represents the different strengths for the topic and the link. Purity, Q and PurQ_β have been calculated for the two mentioned frameworks.

According to Table 2, Modularity and Purity has higher values in the topic-oriented framework, since the basic network is partitioned into topical clusters, and each identified community includes members who have the same topic of interest. Therefore, the topic-oriented community detection framework has a higher value of PurQ_β for all five values of β.

6. Conclusion

This paper evaluates the effect of topic consideration in finding more meaningful communities in social networking sites in

which the users express their feelings toward different objects (like movies) by means of rating. Therefore, the network is partitioned into different topical clusters in which the nodes have the same topic of interest. Then, a community detection algorithm is applied to the topical clusters in order to detect communities. After that, a comparison has been performed between the results of topic-oriented community detection and the results of classical community detection in which no content analysis is performed. The experimental results indicate that the results of topic-oriented community detection will be improved when it is joined with topic analysis.

There is a plenty of room to study the community detection problem in real complex networks which contain huge amounts of information of different natures. Therefore, in future works we have a plan to work on the effect of other kinds of contents in the network, like the communications' content analysis, in finding more meaningful communities in social networking sites in which the users express their feelings toward different objects with rating.

References

Arenas, A., Duch, J., Fernández, A., Gómez, S., 2007. Size reduction of complex networks preserving modularity. New J. Phys. 9, 176.

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

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008.

Chakraborty, A., Ghosh, S., Ganguly, N., 2012. Detecting overlapping communities in folksonomies. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. Publishing, pp. 213–218.

Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A., 1990. Indexing by latent semantic analysis. JASIS 41, 391–407.

Ding, Y., 2011. Community detection: topological vs. topical. J. Informetr. 5, 498–514.

Dooms, S., De Pessemier, T., Martens, L., 2013. MovieTweetings: a Movie Rating Dataset Collected From Twitter. In: Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys'13.

Girvan, M., Newman, M.E., 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99, 7821–7826.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Publishing, pp. 50–57.

Jing, L., Ng, M., Huang, J., 2007. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. IEEE Trans. Knowl. Data Eng. 19, 1026–1041.

Lancichinetti, A., Fortunato, S., 2012. In: Consensus clustering in complex networks. Sci. Rep. 2.

Leicht, E.A., Newman, M.E., 2008. Community structure in directed networks. Phys. Rev. Lett. 100, 118703.

Leskovec, J., Lang, K.J., Mahoney, M., 2010. Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International Conference on World Wide Web. Publishing, pp. 631–640.

Newman, M.E., 2004. Analysis of weighted networks. Phys. Rev. E 70, 056131.

Newman, M., 2011. Communities, modules and large-scale structure in networks. Nat. Phys. 8, 25–31.

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

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

Tang, J., Gao, H., Liu, H., 2012. mTrust: discerning multi-faceted trust in a connected world. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. Publishing, pp. 93–102.

Tang, J., Gao, H., Liu, H., Das Sarma, A., 2012. eTrust: understanding trust evolution in an online world. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Publishing, pp. 253–261.

Tang, J., Gao, H., Hu, X., Liu, H., 2013. Exploiting homophily effect for trust prediction. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining. Publishing, pp. 53–62.

Tang, J., Hu, X., Gao, H., Liu, H., 2013. Exploiting local and global social context for recommendation. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Publishing, pp. 2712–2718.

Wang, D., Kwon, K., Sohn, J., Joo, B.-G., Chung, I.-J., 2014. Community topical "fingerprint" analysis based on social semantic networks. In: Advanced Technologies, Embedded and Multimedia for Human-centric Computing. Publishing, pp. 83–91.

Zhao, A., Ma, Y., 2012. A semantically structured approach to service community discovery, semantics, knowledge and grids (SKG). In: 2012 Eighth International Conference on. Publishing, pp. 136–142.

Zhao, Z., Feng, S., Wang, Q., Huang, J.Z., Williams, G.J., Fan, J., 2012. Topic oriented community detection through social objects and link analysis in social networks. Knowl. Based Syst. 26, 164–173.

Zhu, Y., Yan, X., Getoor, L., Moore, C. 2013. Scalable text and link analysis with mixed-topic link models. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Publishing, pp. 473–481.

Ziegler, C.-N., McNee, S.M., Konstan, J.A., Lausen, G., 2005. Improving recommendation lists through topic diversification. In: Proceedings of the 14th International Conference on World Wide Web. Publishing, pp. 22–32.