Decision Support Systems 54 (2013) 870–879

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

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

User community discovery from multi-relational networks

Zhongfeng Zhang [a,1], Qiudan Li [a,*], Daniel Zeng [a,b,2], Heng Gao [a,1]

[a] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Department of Management Information Systems, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Article history:
Received 18 July 2010
Received in revised form 2 April 2012
Accepted 18 September 2012
Available online 28 September 2012

Keywords:
Community discovery
Multi-relational network
Author topic model
Non-negative matrix factorization

ABSTRACT

Online social network services (SNS) have been experiencing rapid growth in recent years. SNS enable users to identify other users with common interests, exchange their opinions, and establish forums for communication, and so on. Discovering densely connected user communities from social networks has become one of the major challenges, to help understand the structural properties of SNS and improve user-oriented services such as identification of influential users and automated recommendations. Previous work on community discovery has treated user friendship networks and user-generated contents separately. We hypothesize that these two types of information can be fruitfully integrated and propose a unified framework for user community discovery in online social networks. This framework combines the author-topic (AT) model with user friendship network analysis. We empirically show that this approach is capable of discovering interesting user communities using two real-world datasets.

1. Introduction

Web 2.0 technology has enabled massive online social networks and made sharing of user-generated contents easy and almost costless. Two-thirds of Americans now use Facebook, Twitter, Myspace, and other social media sites; and 43% are visiting these sites more than once a day.[3] By May 2010, social networks have become more popular than search engines in U.K., accounting for 11.88% of all U.K. Internet visits.[4] Usually, a social network involves multiple types of relations among different social actors. For instance, on Twitter, a user can specify whom to follow to construct an explicit friendship network. At the same time, this user's posted tweets provide important clues about her interests and such interests across the user community can be used to derive implicit "similarity" relationships among these users. The network embedding multiple types of relations, either explicit or implicit, is called a multi-relational network. Studying multiple relationships is gaining momentum in the literature recently. A case study of homophily on LiveJournal [1] has shown that users' friendships and interests are strongly interlinked. Hernandez [4] introduced some social network principles in online communities. Researchers have also investigated how to combine the friendship network and user-generated contents to cluster users [16,29].

User community changes the way people communicate and affects social interaction [37,38]. Discovering user communities may assist the setup of efficient recommender systems for targeted marketing, improving the quality of social information retrieval, among others [3,25–27,30,40]. For instance, Nie et al. [30] utilized the relevance of communities to improve web page ranking. Online user communities have also emerged as a thriving force in e-Commerce [2]. Spaulding [25] explained how firms could successfully interact with user communities using social contract and trust theory. Ganley et al. [4] examine a popular website Slashdot to test users' social network structure, which would potentially increase the opportunities for monetization. Chiu et al. [27] investigate people's knowledge sharing behavior in virtual communities to help identifying their motivations in communities. In [3], the authors investigated how consumers take advantage of virtual communities as social and information networks, and how this influences their decision making. The identified user communities can also help understand the structural properties of the social network and find the influential users about certain topics, which in turn will help users locate the latent friends they may be interested in.

Most prior work on user community detection has focused on analyzing either user friendship networks [6–8,17] or user-generated contents [12,14] but not both at the same time. The former techniques usually ignore the content generated by users. However, intuitively, two users who have posted similar contents might share common interests and join the same communities, even if no explicit friendship connection exists between them. On the other hand, the latter strategies do not take the friendship connections among users into consideration. Such explicit friendship networks can provide important clues to community discovery.

Being friends "makes a pair of users more likely to share common interests" [1]. In the study of recommender systems, it was shown

* Corresponding author. Tel.: +86 10 62558794.
E-mail addresses: zhongfeng.zhang@ia.ac.cn (Z. Zhang), qiudan.li@ia.ac.cn (Q. Li), zeng@email.arizona.edu (D. Zeng), heng.gao@ia.ac.cn (H. Gao).
[1] Tel.: +86 10 62636334.
[2] Tel.: +1 520 621 4614.
[3] http://www.socialnetworkingwatch.com/2010/06/social-media-up-230-since-2007.html
[4] http://eu.techcrunch.com/2010/06/08/report-social-networks-overtake-search-engines-in-uk-should-google-be-worried/

0167-9236/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2012.09.012

that more accurate recommendations could be made by taking into account both friendship networks and user-generated contents [31]. Similar strategies were also proven to be effective in document retrieval [32] and document classification [33]. This research stream suggests the practical value of a multi-relational approach. In the context of community discovery, work on multi-relational approaches is recently emerging (e.g., [16,28,29]).

This paper focuses on the problem of discovering user communities from multi-relational networks of SNS. We present a unified framework, which combines the author-topic (AT) model with social network analysis (SNA). The AT model, which deals with user-generated content information, is a domain sensitive model, while the SNA methods focus on user friendship networks. Users in the community identified with our approach have dense friendship connections as well as share common content interest. The efficacy of the proposed framework is evaluated using two real-world social network datasets, one from Delicious, a popular social bookmarking site, and the other from Twitter, the most popular microblogging site. Empirical analyses have shown that our algorithm could discover meaningful communities and the topics discussed by these communities in a unified way. Compared to the state-of-the-art, our new framework has resulted in comprehensive performance of closer friendship and higher content interest similarity in the extracted communities.

The rest of the paper is organized as follows. The literature review is presented in Section 2. Section 3 presents in detail the problem definition and our framework. Section 4 introduces the detailed algorithm. The empirical analysis is conducted in Section 5. Finally, Section 6 concludes the paper with a summary and discussion of the future work.

2. Literature review

Most user community detection methods fall into two categories, the network-based and the content-based. The network-based methods construct network structures among users and then split the network into different sub-networks. Technique-wise, these methods are based on graph partitioning in graph theory. The content-based methods discover users with common interests by analyzing the similarity between these users' posted contents. In this section, we review both types of methods.

2.1. Network-based methods

In many social networks, individuals form communities by specifying and establishing friendship connections with each other. The network-based methods aim to find communities such that the friendship connections are dense within communities and sparse between them. Traditional graph partition methods, such as degree-based methods and max-flow min-cut methods, are used to divide the network into groups of predefined size, such that the number of connections lying between the groups is minimal [5]. Spectral clustering techniques partition the network into clusters using the eigenvectors of its related matrices (e.g., Laplacian matrix) [5]. The GN [6] method selects links among users according to edge centrality. Palla et al. [8] proposed a clique percolation method (CPM) based on the concept that the internal links of a community are likely to form cliques due to their high density. Shen et al. [9] proposed to identify overlapping community structures from the maximal clique network of the original network, using modularity optimization methods. Evans et al. [10] introduced a link partition approach for overlapping community structure discovery. Lee and Seung [18] firstly investigated the algorithm of non-negative matrix factorization and it became widely used soon afterwards. Zhang et al. [11] proposed to discover fuzzy community structures in complex networks based on non-negative matrix factorization (NMF). The complex network theory was applied to analyze open-source software systems and structural properties of social interaction in collaborative tagging systems, respectively [41,42].

2.2. Content-based methods

The content-based methods link users and their posted contents via latent topics. Users interested in the same topic are grouped into a community. Steyvers et al. [12] proposed the author-topic (AT) model to explore the relationships among users, documents, topics, and words. It represents a topic as a multinomial distribution over words and models a user as probability distribution over different topics. McCallum et al. [13] presented the author-recipient-topic (ART) model to discover users with similar topic interests, which conditions the topic distribution on the sender–recipient relationships. Based on the ART model, Pathak et al. [15] introduced a community-author-recipient-topic (CART) for community extraction from the Eron email corpus, by leveraging both topic and document link information from the social network. Peng et al. [43] proposed a unified user profiling scheme which makes good use of all types of co-occurrence information in the tagging data. Relying on people's information in database, [39] developed an intelligent secretary agent system to help arrange efficient meetings among people who share similar interests. These models often ignore the explicit friendship connections among users, and may not properly predict users' community memberships.

In this work, we propose a new framework which utilizes both friendship networks and content analysis to discover user communities. At the core of this framework is the NMF-AT algorithm, which performs matrix factorization on the friendship network and author-topic analysis on the user-generated contents. One research closely related to ours combines topic modeling with network regularization [16]. In their work, pLSA was adopted for topic extraction and the graph harmonic function was used for community analysis. Finally, the topical community was extracted by performing topic mapping. However, the authorship of contents was not taken into consideration during the topic extraction procedure in [16]. Instead of performing topic mapping, we tend to extract community topics directly with author-topic analysis.

3. Problem definition

In this section, we first define the terminologies related to community discovery. We then present the framework of our approach for discovering user communities from multi-relational networks.

3.1. Terminology definition

Fig. 1(a) and (b) show examples of multi-relational networks from Twitter and Delicious, respectively. In Twitter, each user is called a twitterer, who can post tweets with a limit of 140 characters, or reply tweets posted by her friends. Each twitterer could follow any other twitterer she is interested in without securing permission. Conversely, she may also be followed by other twitterers. On Delicious, a user could bookmark a url with his own tags, and add interested users into his network to make friends. The friendship networks are marked with dashed lines in Fig. 1.

Given a multi-relational network as shown in Fig. 1, we represent it as a graph $G=(V, E)$, where V is the set of actors in the network, and E is a set of edges indicating the connections among actors in V. For instance, in Fig. 1(a), $V=\{U, T, R, W\}=\{<U1,\ldots,U5>, <T1,\ldots,T4>, <R1,R2>, <w1,w2, w3>\}$, where Ui $(i=1,\ldots,5)$ represents a twitterer, Tj $(j=1,\ldots,4)$ a tweet, wk $(k=1,2,4)$ a word in the vocabulary, and $R1/R2$ reply to other tweets (which is also called a tweet). $E=\{<U1,U2>, \ldots, <U1,T1>, \ldots, <T1, w1>, \ldots,\}$ indicates the relationship among twitterers, tweets, and words. The edge $<U1, U2>$ indicates that twitterer $U1$ has followed $U2$. The edge $<U1,T1>$ implies the twitterer $U1$ has posted tweet T1. The edge $<T1,w1>$ indicates that tweet $T1$ is composed of word $w1$.

The friendship network associated with G is a subgraph $F=<U, Eu>$, where $Eu \subset E$ is a set of edges among users. In the twitter case, $Eu=\{<U1, U2>, \ldots, <U5,U4>\}$. The fact that a twitterer $U1$ follows $U2$ does not necessarily imply that $U2$ has followed $U1$. In such a