Neurocomputing 179 (2016) 219–227

Contents lists available at ScienceDirect

Neurocomputing

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

An improved density peaks-based clustering method for social circle discovery in social networks

Mengmeng Wang [a,b], Wanli Zuo [a,b], Ying Wang [a,b,c,*]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, PR China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun 130012, PR China
[c] College of Mathematics, Jilin University, Changchun 130012, PR China

ARTICLE INFO

Article history:
Received 18 May 2015
Received in revised form
6 November 2015
Accepted 29 November 2015
Communicated by Yongdong Zhang
Available online 15 December 2015

Keywords:
Discovering overlapping social circles
Improved density peaks-based clustering method
In-link Salton metric
Out-link Salton metric
Social networks

ABSTRACT

With the development of Internet, social networks have become important platforms which allow users to follow streams of posts generated by their friends and acquaintances. Through mining a collection of nodes with similarities, community detection can make us understand the characteristics of complex network deeply. Therefore, community detection has attracted increasing attention in recent years. Since the problem of discovering social circles is posed as a community detecting problem, hence, in this paper, targeted at on-line social networks, we investigate how to exploit user's profile and topological structure information in social circle discovery. Firstly, according to directionality of linkages, we put forward in-link Salton metric and out-link Salton metric to measure user's topological structure. Then we propose an improved density peaks-based clustering method and deploy it to discover social circles with overlap on account of user's profile- and topological structure-based features. Experiments on real-world dataset demonstrate the effectiveness of the proposed framework. Further experiments are conducted to understand the importance of different parameters and different features in social circle discovery.

1. Introduction

With the advent of online social networks, social network analysis has gradually become a hot issue in both academia and industry. Since community structure, which often represents specific organized groups of users with similar attributes, hobbies or closer relationships [1], is a significant property of social networks, so community detection is one of the basic research problems in social network analysis. Detecting community is very important for understanding the characteristics of complex network, discovering latent topology, predicting network evolution and so on [2]. Besides, identifying community structure can facilitate many tasks such as following/follower recommendation [3], task allocation [4], proximity alignment [5], maximizing influence [6], retweeting behavior prediction [7], mining cybercriminal networks [8] and so forth. A novel function has been provided in some major social networks: users can categorize their friends into social circles which can be used to filter status updates posted by distant acquaintances, hide personal information from coworkers and share groups of users that others may wish to follow [9]. Therefore, social circle discovery which can fall into the domain of community detection has attracted increasing attention in recent years. As stated, our work on social circle discovery is motivated by its broad application prospect.

The purpose of community detection is to find a group of users with similar ideas, beliefs, motivations or other common features so as to better understand social networks. However, most of existing community detection approaches which only considered structural features (e.g., links) [10] may ignore much information that associated with community, such as user's background information and interaction information [11]. Besides, isolating user's relationship with user's contents may result in finding unreasonable community structure, while most community discovery algorithms which considered both types of information were usually complex. Hence, in this paper, we propose an improved density peaks-based clustering method which incorporates both structural and attribute information of users for social circle discovery in social networks (denoted as DPSCD), and our main contributions are summarized next.

(1) Put forward in-link Salton metric and out-link Salton metric according to directionality of linkages to achieve a better representation on adjacent degree between users in directed social networks;

(2) Improve a fast clustering method with novel density estimator and extra social circle integration step in order to better

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun 130012, PR China. Tel.: +86 13604413931.
E-mail addresses: wmmwwlh@126.com (M. Wang), zuowl@jlu.edu.cn (W. Zuo), wangying2010@jlu.edu.cn (Y. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.11.091
0925-2312/

adapt to large statistical errors, followed by employing it to detect overlapping social circles in social networks; and

(3) Evaluate DPSCD on real-world dataset Facebook, Google+, Twitter and elaborate the importance of different parameters and different features on social circle discovery.

The rest of paper is organized as follows: Section 2 describes the related work; Section 3 defines the method we propose; Details of the experimental results and dataset which is used in this study are given in Section 4. Finally conclusion appears in Section 5.

2. Related work

It has been pointed out that social circle discovery can be formulated as a community detection problem on a user's ego network (the network of friendships between his/her friends) [9–12]. A great deal of works have been done on community detection, Newman and Girvan [13] first defined a community as a subgraph containing nodes which were more densely linked to each other within such subgraphs and sparse between them. Community discovery provides an important means to understand the structure of complex networks deeply [14], therefore, the issue of detecting community in social networks has received increasing attention in recent years. Existing community detection algorithms can be roughly categorized into three groups: relationship-based method, content-based method and comprehensive method where users' relationships and contents are merged with users' attributes.

Since topological structure can influence individuals' behaviors in the network [15], as a consequent, discovering communities based on users' relationships is still a mainstream method. In relationship-based algorithms, through utilizing properties of relationships, communities with closer internal relationships are detected. Li et al. [16] first proposed two new algorithms based on evolutionary algorithm and clonal selection algorithm (denoted as EA-SN and CSA-SN, respectively). Then, they integrated a hill climbing (HC) strategy into EA-SN and CSA-SN to form two new memetic algorithms (EAHC-SN and CSAHC-SN). The experimental results not only showed the capability and high efficiency of EAHC-SN and CSAHC-SN in successfully detecting communities from signed networks, but also indicated that both the two objective functions (improved modularity and improved modularity density) were efficient to some extent. Ma et al. [17] presented a seed insensitive method for local community detection which estimated similarity among vertices by investigating vertices' neighborhoods and revealed a local community by maximizing its internal similarity and minimizing its external similarity simultaneously. Extensive experimental results on both synthetic and real-world data sets verified the effectiveness of the proposed algorithm. Cai et al. [18] proposed a novel discrete PSO algorithm for identifying community structures in signed networks. In order to make PSO be proper for discrete scenarios, they redesigned particles' status in a discrete form, followed by reformulating particles' updating rules through making use of topological structure of signed network. Extensive experiments demonstrated that the proposed method was effective and promising. Li et al. [19] first improved nonnegative matrix factorization method with modeling network as a weighted directed graph and using diagonally dominant matrix as constraint condition to obtain community membership of each node as well as interactions between communities. The results demonstrated that the proposed method was useful and applicable both in weighted directed model and undirected model for community discovery over other related matrix factorization methods. Rhouma et al. [20] proposed an overlapping community detecting algorithm called DOCNet (Detecting overlapping communities in Networks) which was based on local optimization of a fitness function and a fuzzy belonging degree of different nodes. The main strategy of this algorithm was to find an initial core and add suitable nodes to expand it until a stopping criterion was met. Experimental results demonstrated that DOCNet was efficient and highly reliable for detecting overlapping groups, compared with four newly known proposals. However, DOCNet model cannot adapt to weighted and directed networks. Qiu et al. [21] first generated a probability transition matrix by applying random walk to a social network, followed by training a Gaussian mixture model using the matrix. And then, overlapping communities were derived by analyzing mean vectors of the Gaussian mixture model. The experiments conducted on synthetic and real dataset demonstrated the feasibility and applicability of the proposed algorithm. Instead of using eigenvectors in spectral clustering algorithms, Huang et al. [22] put forward a regularized spectral clustering algorithm which chose sample matrix of social network to construct a target function that can partition social network naturally. The experiments shown the proposed method achieved good results with relatively smaller computational cost compared to spectral clustering algorithm. Wu et al. [23] introduced a cosine-pattern-based community extraction framework. It first extracted the so-called asymptotically equivalent structures (AESs) from networks, from which nodes were further partitioned into crisp communities using any of existing methods. A novel cosine-pattern mining algorithm based on the ordered anti monotone of cosine similarity was thus proposed for the efficient extraction of AESs. Experiments on various real-world social networks demonstrated the advantage of extracting view of community detection.

However, Dang and Viennet [24] pointed out that in real-world networks, in addition to topological structure (i.e., links), content information was also available. Besides, considering network structural information only may fail to detect interpretable overlapping communities since structural information of online social networks is often sparse and weak. Sang and Xu [25] held the view that the social links were well recognized forces that govern the behaviors of involved users as well as the dynamics of social networks. Additionally, through splitting all the Flickr user pairs into two parts, i.e., with relations and without relations, respectively and calculating the average of the common contact number, common interested group number and tag-based similarity in the respective user pairs, Yan et al. [26] pointed out that users generally had more common contacts and common interested groups with their friends than other people. Besides, friends may also use more similar tags in their uploaded images which may be influenced by each other. So some scholars began to calculate distance/similarity between users on their generated information.

Yin et al. [27] incorporated community discovery into topic analysis in text-associated graphs and proposed a community-based topic analysis framework called LCTA (Latent Community Topic Analysis). The proposed framework handled both topic modeling and community discovery to guarantee the topical coherence in the communities so that users in the same community were closely linked to each other and shared common latent topics. They compared different methods and performed extensive experiments on two real datasets. The results confirmed the hypothesis that topics could help understand community structure, while community structure could help model topics. Taking Flickr as one exemplary social media platform, Zhuang et al. [28] found that the taste/interest of a user in photos can be implicitly mined from the photos uploaded by the user and they proposed a content-aware low-rank matrix recovery technique for community discovery. First, they modeled the observed indicator matrix of the Flickr community as a summation of a low-rank true matrix and a sparse error matrix. And then, they formulated an optimization