Expert Systems With Applications 58 (2016) 10–19

Contents lists available at ScienceDirect

Expert Systems With Applications

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

An adaptive random walk sampling method on dynamic community detection

Yu Xin [a,*], Zhi-Qiang Xie [a], Jing Yang [b]

[a] College of Computer Science and Technology, Harbin University of Science and Technology, Heilongjiang, 150001, China
[b] College of Computer Science and Technology, Harbin Engineering University, Heilongjiang, 150001, China

ARTICLE INFO

Article history:
Received 17 March 2015
Revised 21 March 2016
Accepted 22 March 2016
Available online 1 April 2016

Keywords:
Random walk
Dynamic community
Community detection

ABSTRACT

With the change of lifestyle and interests, the people's social activities have a dynamic changing tendency. Therefore, the static community could not reflect the real activities. For the 'community' in the social network is the aggregate of people's activities, thus the dynamic community could be detected by simulating the individual freewill. The individual tends to get in touch with the closest friends. By that a direction from one node to its closest nodes can be obtained, and the formed directed network could easily find out the communities. It is different from the traditional community detection policies, which only consider the global topological structure of the social network. Accord to the theory above, we designed the RWS (Random Walk Sampling) method to detect the overlapping communities, utilizing the random walk method to find the closest friends for each node. As the topological structure changing, the proposed ARWS (Adaptive Random Walk Sampling) could make the impacted nodes find out the new closest friends and the changed communities adaptively. The ARWS only update the impacted nodes and communities as the dynamic events occurring, while the traditional dynamic community detection methods need to break up and restructure the communities after the topology changing, because the tradition methods are based on the global topological structure. Therefore, the ARWS has a lower cost than the traditional methods. Furthermore, the ARWS focus on the individual, fitting to the decentralized computing framework, such as distributed computation and cloud computing. That is the trend of the artificial intelligence.

1. Introduction

The people's social activities are variable, and the habits and hobbies are slowly changing, giving arise to the evolution of social network. Taking the case of social network sites, there exist many new registered users and disappear users, and the relationships of users appear to be dynamic, such as the establishment of new friendships and the blanking of old friendships. Kumar, Novak, and Tomkins (2010) designed three policies to simulate the evolution of social network. In the generative process, the users in social network are classified into three groups: passive, linker and inviters. The passive users join the network out of curiosity or at the insistence of a friend, but never engage in any significant activity. Inviters are interested in migrating an offline community into an online social network, and actively recruit their friends to participate. Linkers are full participants in the growth of the online social network, and actively connect themselves to other members. Various dynamic social network structures, such as Flickr and Yahoo!360, can be simulated by adjusting the mixture coefficient. Therefore, the research on the evolution of social networks is the practical needs on social computing, much research on community evolving (Cuzzocrea & Folino, 2013a; Cuzzocrea, Folino, & Pizzuti, 2013b; Konstantinidis, Papadopoulos, & Kompatsiaris, 2013), community tracing (Cuzzocrea & Folino, 2013a; Cuzzocrea et al., 2013b; Zhong, An, Gao, & Sun, 2014) and dynamic simulation (Takaffoli, Fagnan, Sangi, & Zaıane, 2011a; Takaffoli, Sangi, Fagnan, & Zäıane, 2011b) has been proposed. These methods evaluate the community evolution in terms of the communities' variations , such as form, dissolve, survive, split and merge. Compared with the existing methods, we have considered the following two objectives. On one hand, need to guarantee the detected communities have a higher rationality (a higher modularity). On the other hand, need to guarantee the communities have a dynamic adaptability to deal with the dynamic events. For that, we proposed the ARWS to deal with the two problems above. Our main contributions in this paper can be summarized as the following:

* Corresponding author. Tel.: +8613836014968.
E-mail addresses: xinyu@hrbeu.edu.cn (Y. Xin), xiezhiqiang@hrbust.edu.cn (Z.-Q. Xie), yangjing@hrbeu.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.eswa.2016.03.033
0957-4174/

(1) The proposed RWS utilize the random walk as the local topological environment analyzing method, to find out the closest friends for the individual. By which, the global network could be simplified into the directed network, only containing the closest relationships. The directions of the nodes are the community clustering directions. By which, the global network could be clustered into the overlapping communities. The process only focuses on the individual, regardless of the global topological structure, which could reduce the complex of the community detection.
(2) The ARWS is designed based on the RWS. The ARWS only update the closest friends of the nodes impacted by the dynamic events. Consequently, the impacted communities could be adjusted adaptively, by accepting or removing the impacted nodes. Compared with the traditional dynamic community detection methods needing to break up and restructure the communities after the dynamic events, the adaptive adjustment of the communities has a much lower cost.
(3) The proposed RWS and ARWS detect the clustering directions of the nodes by analyzing the local topological environment. Therefore, the ARWS focusing on the individual has a decentralized computing manner, which could be applied into the distribution computing environment.

Furthermore, the drawbacks of ARWS are the following:

(1) The ARWS has too much input parameters, such as the sampling frequency *s*, the maximum walking step *r*, the overlapping threshold *h*, the favorable impact threshold *e*, and could not find out the accurate values for the parameters. That would lead to the inaccurate results.
(2) The random walk method used by ARWS is homogeneous random process, implying the walking does not consider the impact of network characteristics on the walker, at each walking step, such as the weight, density, similarity.

2. Related work

Currently, dynamic community detection methods can be classified into four categories: dynamic clustering, objective function optimization, representative node (community) detection, dynamic probability modeling.

(1) Dynamic clustering methods. These methods are mainly based on the EC (Evolutionary Clustering) proposed by Chakrabarti, Kumar, and Tomkins (2006), seeing the time snapshot as clustering sample unit, integrated modeling the node distribution in time snapshot *t* and *t*-1. Chi, Song, Zhou, Hino, and Tseng (2007) proposed the PCQ (Preserving Cluster Quality) and (Preserving Cluster Membership) methods based on EC, where the PCQ emphasis on the cluster density and PCM on the node similarity in the same cluster. Lin proposed the FacetNet (Lin, Chi, Zhu, Sundaram, & Tseng, 2008) method based on EC, which established the snapshot cost function according to the community distribution in time snapshot *t*. Kim proposed the PDEM(Particle and Density based Evolutionary Clustering) (Kim & Han, 2009) improving the FacetNet, which took into account of the drawback of Facetnet on needing to preset the number of communities and not being allowed to change it.
(2) Objective function optimization methods. These methods estimated the community evolution by the structure changes, the process of which is guided by the optimization of community density function or modularity. Blondel, Guillaume, Lambiotte, and Lefebvre (2008) applied the FN Newman (2004) into the dynamic community detection, which is the combination of local optimizing and hierarchical clustering. Dinh, Xuan, and Thai (2009) suggested the MIEN(Modules Identification in Evolving Networks), which employed the incremental modularity optimizing method, compressing the network into several representative communities. Nguyen proposed the AFOCS (Adaptive Finding Overlapping Community Structure) (Nguyen, Dinh, Tokala, & Thai, 2011a; Nguyen, Dinh, Xuan, & Thai, 2011b) and QCA(Quick Community Adaption) methods. These two methods designed four adjusting policies specific to the four dynamic events (node adding, node deleting, link adding, link deleting), and improved the community density function and modularity respectively, allowing for the dynamic community detection. Guo, Wang, and Zhang (2014) proposed the ECSD(Evolutionary Community Structure Discovery) aiming at weighted dynamic community detection. In the process of ECSD, as the dynamic events occurrence, the increment of merged communities is treated as the criterion on community merging decision.
(3) Representative node (community) detection methods. These methods facilitated the adjusting process via tracing the representative nodes or communities. Zhang, Wang, Wang, and Zhou (2009) proposed the BSP (Bulk Synchronous Parallel), which established the similarity measurement using the number of common neighbors and the connections. Duan, Li, Jin, and Lu (2009) suggested the stream-group, which used the transition probability matrix to establish the relevance measurement between nodes, and proposed the compactness model to evaluate the local community tightness. Bourqui et al. (2009) measure the similarity of communities between the adjacent time snapshots, and merged the communities with a high similarity. Chen, Wilson, Jin, Hendrix, and Samatova (2010) proposed the representative-based community evolution tracing method, which made an assessment on the detected communities to find out the stable communities. When the network structure varies, it only traces the stable communities, by which the comparing times can be reduced. Takaffoli et al. (2011a, 2011b) gave an evolution analyzing method specific to the nodes, which adopts the steady-state model and impact model to recognize the stable nodes and representative nodes. Duan, Li, Li, and Lu (2012) gave an incremental *K*-clique community adjusting method utilizing the dynamic updating policy of DFS tree. Ma and Huang (2013) proposed the CUT(Community Update and Tracking) model, which detected the seed community (namely clique community) as the representative community. When the dynamic events occurrence, the CUT only updated the affected seed community to reduce the community detection cost.
(4) Dynamic probability modeling methods. These methods assume the community distributions in each time snapshot as the samples of latent community distribution, and establish the probability model to associate the samples with latent community. Sarkar and Moore (2005) adopted the kernel method to construct the probability relation between nodes, then utilized probability relation between nodes and communities in each time snapshot to construct the Bayesian model between nodes and latent communities. Sun, Faloutsos, Papadimitriou, and Yu (2007) proposed the DPM (Dirichlet Process Mixture Model) which sampled the nodes in each time snapshot, constructing the Bayesian model between nodes in the adjacent time snapshot, thus the latent communities can be detected by LDA method.

In addition to the above methods, Sun et al. (2007) proposed the GraphScope method, adopting the coding optimization.