Information Sciences 388–389 (2017) 37–47

Contents lists available at ScienceDirect

Information Sciences

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

Fast graph clustering with a new description model for community detection

Liang Bai[a,b,c,*], Xueqi Cheng[b], Jiye Liang[a], Yike Guo[c]

[a] School of Computer and Information Technology, Shanxi University, Taiyuan, 030006, Shanxi, China
[b] Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
[c] Department of Computing, Imperial College London, SW7, London, United Kingdom

ARTICLE INFO

Article history:
Received 13 October 2015
Revised 16 December 2016
Accepted 7 January 2017
Available online 9 January 2017

Keywords:
Graph clustering
Community detection
Community description model
Evaluation criterion
Iterative algorithm

ABSTRACT

Efficiently describing and discovering communities in a network is an important research concept for graph clustering. In the paper, we present a community description model that evaluates the local importance of a node in a community and its importance concentration in all communities to reflect its representability to the community. Based on the description model, we propose a new evaluation criterion and an iterative search algorithm for community detection (ISCD). The new algorithm can quickly discover communities in a large-scale network, due to the average linear-time complexity with the number of edges. Furthermore, we provide an initial method of input parameters including the number of communities and the initial partition before algorithm implementation, which can enhance the local-search quality of the iterative algorithm. The proposed algorithm with the initial method is called ISCD+. Finally, we compare the effectiveness and efficiency of the ISCD+ algorithm with six representative algorithms on several real network data sets. The experimental results illustrate that the proposed algorithm is suitable to address large-scale networks.

1. Introduction

Cluster analysis is a branch in statistical multivariate analysis and unsupervised machine learning. The goal of clustering is to group a set of objects into clusters so that the objects in the same cluster have high similarity but are very dissimilar with objects in other clusters [14]. To solve this problem, various types of clustering algorithms, such as partitional clustering and hierarchical clustering, have been proposed in the literature (e.g., [18] and references therein). Recently, increasing attention has been paid to analyzing cluster structures in complex networks since the data are modeled as networks in many complex systems [38], e.g., social networks and biological networks. In network analysis, cluster structure is also called "community structure" [10,30] which has been shown to be an important property of networks. Intuitively, a community (cluster) in a network consists of a cohesive group of nodes that are relatively densely connected to each other but sparsely connected to other dense groups. Community detection aims to identify the communities by only using the information encoded in the network topology. It can be seen as a procedure of *graph clustering*. Community detection becomes one of the most important tasks to explore and understand how the networks work [11].

* Corresponding author.
E-mail addresses: sxbailiang@hotmail.com (L. Bai), cxq@ict.ac.cn (X. Cheng), ljy@sxu.edu.cn (J. Liang), y.guo@imperial.ac.uk (Y. Guo).

http://dx.doi.org/10.1016/j.ins.2017.01.026
0020-0255/

To resolve the community detection problem, various graph clustering approaches have been developed, including latent space models, non-negative matrix factorization, block model approximation, spectral clustering, label propagation, and modularity maximization. According to applications for different scientific needs, these models have different definitions of communities or clustering criteria [40]. Latent space models [36] mainly map nodes of a network into a low-dimensional Euclidean space. The proximity between the network connectivity nodes is kept in the new space; then, the nodes are clustered in the low-dimensional space by using traditional clustering algorithms such as *k*-means [24] and linkage [42]. Like the latent space models, non-negative matrix factorization models [22,41,44] transfer the adjacency matrix of a network into a low-dimensional matrix, then cluster it by *k*-means or linkage. Block model approximations see a community detection problem as a matrix blocking problem, which reorder the index of each node according to their community membership and approximate a given network by a block structure [7]. Each block represents a community. Spectral clustering models [13,37] view the community detection as a graph partitioning, which apply spectral analysis to obtain the cut minimization. Label propagation models mainly use the neighbor information of each node to determine its label and do not need any prior knowledge of community structure. The representative algorithm of LPA was proposed by Raghavan et al. [32]. It has greatly received attention for its nearly linear time complexity in finding communities. However, since the label of each node depends on those of other nodes, the algorithm can only linearly propagate the labels. In addition, the convergence speed and clustering effectiveness of the algorithm are very sensitive to the update order of label information. Therefore, several improved LPA algorithms are developed in [2,16,39]. Modularity maximization models [8,9,11,26,29] transform a community detection problem into a modularity maximization problem. Modularity is a commonly used criterion for community detection, which measures the strength of a community partition for real-world networks by taking into account the degree distribution of nodes. The type of the algorithms mainly apply different hierarchical clustering strategies to partition networks, which is very time-consuming. The fast unfolding algorithm proposed by Blondel et al. [5] is a fast heuristic method for the modularity optimization. The algorithm uses the idea of the label propagation models to reduce the computing cost. Compared to other algorithms for modularity maximization, the fast unfolding algorithm has good scalability for large networks. Additionally, Rosvall and Bergstrom develop an information-theoretic model for community structure [34]. They transfer the problem of community detection into a information coding problem. Furthermore, an information map algorithm of random walks [35] is proposed to solve the optimization problem. There are several studies about the survey of the performance of the existing community detection algorithms, such as [3,15]. The authors compared the performance of these algorithms in real networks and analyzed the strength and weakness of each algorithm.

Many existing community detection models have been successfully applied to different areas. However, there are two main problems in the clustering process. One problem is that most of these models need expensive computing costs including the transformation of a network into a $n \times k$ matrix or hierarchical clustering strategy. These costs limit their efficiency in dealing with large-scale networks. The other is that there is a lack of an effective community description model which is used to summarize and characterize the community. After obtaining the model, we can quickly determine whether a node belongs to the community based on a similarity or distance measure. This can enhance the expandability of community detection for new input nodes in a network. In cluster analysis, the *k*-prototypes-type algorithms, such as *k*-means, are a kind of partitional clustering technique and well known for efficiently clustering large-scale spherical data sets. They usually use a virtual or real point on a given data set to represent a cluster. Unfortunately, these types of algorithms mainly deal with object-attribute data sets. It is very difficult for a network without the feature space information to compute the center of a community to represent it. Currently, many algorithms are proposed to transform a network into an object-attribute data set and cluster it by the traditional clustering algorithms. However, feature extraction maybe lead to information loss and high transformation costs. Therefore, it is an important issue to directly handle the raw network data. However, few scholars have discussed how to directly describe a community by using nodes in the *k*-prototypes-type clustering.

Motivated by the above idea, we design a new *k*-prototypes-type algorithm for quickly clustering large-scale networks. The major contributions of this paper are as follows:

- Unlike the traditional *k*-prototypes-type algorithms, we propose a community description model, which does not make use of a node but multiple nodes with different weights (representability) to represent the community. The new description model can sufficiently reflect the characteristics of the communities.
- Based on the description model, we provide a community detection criterion. It evaluates the quality of a partition result from two aspects: the external-link separation among communities and the internal-link compactness within communities.
- Based on the evaluation criterion, we propose an iterative search algorithm for community detection (ISCD) which applies a local search to partition a network into *k* communities. The new algorithm inherits the advantage of the *k*-prototypes-type clustering. Namely, the algorithm can address large-scale networks, since its average time complexity is linear with the number of edges.
- Like the traditional *k*-prototypes-type algorithms, the ISCD algorithm needs initial parameters, including initial partition and the number of communities. We propose an initial method for these parameters to enhance the local-search quality of the ISCD algorithm. The new algorithm with the initial method is called ISCD+.

The following is the outline of this paper. Section 2 introduces a community description model. Section 3 presents an evaluation criterion and an iterative algorithm for community detection. Section 4 provides an initial method for parameters