Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 31 (2014) 85 – 94

Procedia Computer Science

2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

Weighted graph clustering for community detection of large social networks

Ruifang Liu[a], Shan Feng[a], Ruisheng Shi[b,c,*], Wenbin Guo[a]

[a]*School of Information and Communication Engineering, Beijing University of Posts and Telecommunications,Beijing 100876, China*
[b]*Education Ministry Key Laboratory of Trustworthy Distributed Computing and Service,, Beijing University of Posts and Telecommunications,Beijing 100876, China*
[c]*School of Humanities, Beijing University of Posts and Telecommunications,Beijing 100876, China*

Abstract

This study mainly focuses on the methodology of weighted graph clustering with the purpose of community detection for large scale networks such as the users' relationship on Internet social networks. Most of the networks in the real world are weighted networks, so we proposed a graph clustering algorithm based on the concept of density and attractiveness for weighted networks, including node weight and edge weight. With deep analysis on the Sina micro-blog user network and Renren social network, we defined the user's core degree as node weight and users' attractiveness as edge weight, experiments of community detection were done with the algorithm, the results verify the effectiveness and reliability of the algorithm. The algorithm is designed to make some breakthrough on the time complexity of Internet community detection algorithm, because the research is for large social networks. And the another advantage is that the method does not require to specify the number of clusters.

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.

Keywords: Micro-blog; User Attractiveness; Community Detection; Weighted Graph Clustering

1. Introduction

Community detection has received significant attention in all kinds of networks,such as the World Wide Web[1], collaboration networks[2,3], biological networks[4], and social networks[5].

With the rapid development of the Internet, it recently have attracted the attention of researches with different algorithms to discover and analyze the potential communities in the Internet. It was known that, the variety of physical social circles, in some levels, could reflect the relationship among people. People in a physical social circle usually also have some contacts in the Internet. Through analyzing community structure of the online social networks, such as FaceBook, Twitter, Sina micro-blog, Renren, we could probably find the potential relationship exists among people. Many applications on Internet, such as recommendation systems, can also benefit from such social network analysis.

* Corresponding author.
E-mail address: shiruisheng@bupt.edu.cn

Selection and peer-review under responsibility of the Organizing Committee of ITQM 2014.
doi:10.1016/j.procs.2014.05.248

Community detection problem has been studied as the graph partitioning problem in computer science for decades and is known to be a NP-hard problem. Many algorithms have been proposed, including hierarchical clustering[5,6,7], random walk based methods[8,9], spectral clustering[10,11], modularity based methods[4,7,12], user profile based methods[13,14]. These methods are all popular methods, but in the real world, most of the networks contain weighted information, however, there are only a few algorithms designed for weighted networks, and most algorithms are difficult to extend to weighted networks unfortunately. Another problem is that many algorithms are not fit for the large-scale networks community detection because of the high computational complexity.

In the paper, we proposed the concept of community attractiveness, with this definition, a clustering algorithm is constructed, named attractiveness-based community detection(ABCD) algorithm, which are introduced in section 3. In section 4, with the analysis on the micro-blog user network, we define the concepts of node weight and edge weight for the network, and present the experimental results, and the performance and execution time compare were done between ABCD algorithm and CNM(Clauset-Newman-Moore) algorithm[12]. Section 5 shows the experimental results on the College Football Team dataset, and section 6 shows the experimental results of a social network called Renren. Conclusions appear in Section 7. The research is for large social networks, and the another advantage is that the method does not require to specify the number of clusters, this number is usually not known in advance and is difficult to estimate.

2. Related Works

GN algorithm[6] is historically important, because it marked the beginning of a new era in the field of community detection, but it requires a time $O(n^3)$ on a sparse graph. CNM algorithm[12] is an improved algorithm, it has essentially linear running time $O(n \log^2 n)$.

Some works were done for Internet social networks. ISCoDe[13] is a framework based on methods for detecting communities over weighted graphs, where graph edge weights are defined based on measures of similarity between individuals interests tag. Slah Alsaleh et al.[14] provide a system using a clustering technique to create sets of communities based on users information, and then similar communities are matched based on users activities. For twitter dataset in a paper[15], twitters are nodes, the count of retweets between A and B is the weight of edge, and then to accentuate clusters with variable density, the experimental result is not good, only a small number of communities were detected in the Twitter dataset.

3. Weighted Graph Clustering

3.1. Problem Statement

Community structure finding can be considered as a graph clustering problem. And this problem can be considered as an optimization problem[16].

We suppose each person or a community has a density value, and each pair of persons or communities has an attractiveness value. The social network is a graph, each person is a node, edges are the relationship between people. Given a sparse graph $G(V, E, W_V, S_E)$ which consists of the node set V, the edge set E, the weight of node set W_V, and the weight of edge set S_E, we are interested in finding the clusters of G as communities.

Undirected graphs are the most common models of networks, where the directions of the connections are unimportant and can be safely ignored. Here we considered only undirected graph. Figure 1 shows such a graph, the weight of node implies the core degree of the person in the network, and the weight of edge means the attractiveness between the two nodes.

The result of graph clustering should partition a graph into several sub-graph(clusters), each part has a weight value, what's more, there are attractiveness values between clusters which similar with the edge weights. The candidate communities should have weights higher than the attractiveness with other clusters.

The optimization objective function is equation (1), $\mathcal{P}$ is the partitions of a graph.

$$\text{argmax}_{\mathcal{P}}\{\sum_{k\in\mathcal{P}} W(k) - \sum_{i,j\in\mathcal{P}} S(i,j)\} \tag{1}$$