Engineering Applications of Artificial Intelligence 25 (2012) 640–647

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

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

A new blockmodeling based hierarchical clustering algorithm for web social networks ☆

Shaojie Qiao [a,*], Tianrui Li [a], Hong Li [a], Jing Peng [b], Hongmei Chen [a]

[a] School of Information Science and Technology, Southwest Jiaotong University, No. 111, Erhuanlu Beiyiduan, Chengdu, Sichuan 610031, China
[b] Department of Science and Technology, Chengdu Municipal Public Security Bureau, No. 136, Wenwu Road, Chengdu, Sichuan 610017, China

ARTICLE INFO

Article history:
Received 18 October 2010
Received in revised form
11 April 2011
Accepted 5 January 2012
Available online 31 January 2012

Keywords:
Web social networks
Hierarchical clustering
Blockmodeling
Structural equivalence
Optimization

ABSTRACT

Cluster analysis for web social networks becomes an important and challenging problem because of the rapid development of the Internet community like YouTube, Facebook and TravelBlog. To accurately partition web social networks, we propose a hierarchical clustering algorithm called HCUBE based on blockmodeling which is particularly suitable for clustering networks with complex link relations. HCUBE uses structural equivalence to compute the similarity among web pages and reduces a large and incoherent network into a set of smaller comprehensible subnetworks. HCUBE is actually a bottom-up agglomerative hierarchical clustering algorithm which uses the inter-connectivity and the closeness of clusters to group structurally equivalent pages in an effective fashion. In addition, we address the preliminaries of the proposed blockmodeling and the theoretical foundations of HCUBE clustering algorithm. In order to improve the efficiency of HCUBE, we optimize it by reducing its time complexity from $O(|V|^2)$ to $O(|V|^2/p)$, where p is a constant representing the number of initial partitions. Finally, we conduct experiments on real data and the results show that HCUBE is effective at partitioning web social networks compared to the Chameleon and k-means algorithms.

1. Introduction

Recently, social network analysis (SNA) has been recognized as a promising and effective technology for studying complex networks, especially for the Internet, which contains a large amount of web pages that grow in the "TB" order of magnitude each day. Because of the characteristic of dynamical structure on the Web, there is an increasing trend that the Web is partitioned into distinct subnetworks. We call such subnetwork as web social network (WSN for short), since it is a social structure made of web pages called "nodes" which are connected by one or more specific types of interdependency, such as similar theme or content related to friendship, kinship, financial exchange, dislike, relationships of beliefs, knowledge or prestige.

SNA is an alternative solution to analyze patterns of relationships and interactions between social actors in order to discover an underlying social structure (Breiger, 2004). It focuses practically on the relationships between social structures and semantic configurations (Pattison, 1994) while the structure of the social network constitutes. There are several SNA techniques that have been applied to analysis the dynamical structure (Strogatz, 2001), discover the essential players (Qiao et al., 2008; Xu and Chen, 2005; Qiao et al., 2011), and predict the evolving trend of a social network (Qiu et al., 2009). But there is rare work focusing on to discover new emerging groups of similar objects in WSNs, which can help electronic dealers or online sailers classify their users by their interactions and interests in order to provide special services for distinct categories of people. How to develop an accurate and efficient clustering algorithm for partitioning WSNs is a challenging problem due to the following characteristics in web social networks:

- The relations among pages in a WSN are very intricate, and the links between pages could have distinct weights, directions and signs. The traditional clustering algorithms including the distance, hierarchical, density based approaches that cannot be directly applied to partition the WSN containing a lot of pages with complex interactions.
- The network structure of the Web could change dynamically over time. For example, on the WWW, pages or links are created and disappeared every minute, which is difficult for us to accurately partition the dynamically evolving network structure.
- Meta-complication (Strogatz, 2001): the various complications can influence each other. For example, the present layout of a power complex network depends on how it has grown over the years—a case where network evolution. The pages could be

☆ This paper is an extended version of a communication given at FLINS 2010.

* Corresponding author. Tel.: +86 13551192302.
E-mail addresses: qiaoshaojie@gmail.com, sjqiao@swjtu.edu.cn (S. Qiao).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2012.01.003

nonlinear dynamically evolved and the state of each page can vary over time in complicated ways. How to group uncertain nodes in a WSN seems difficult because of these complications.

Blockmodeling has been recognized as an analysis tool for large networks with position as a central concept, which seeks to group units that have substantially similar patterns of relations with others and can be used to discover the relations of structure equivalence among distinct clusters (Breiger, 2004). Particularly, the structure equivalence emerged as a basic concept for analyzing the social network representations of social structure (Burt, 1988). As addressed previously, WSN has complex relations among social actors (i.e., web pages) and the importance of actors is defined beyond the position relation among them. So, it is appropriate to use the block models to partition the social relations among web pages into disjoint classes.

From the above discussions, we need to develop an effective clustering algorithm in order to analyze complex networks in an efficient fashion, especially for WSNs. To achieve these goals, we made the following contributions in this study:

1. In order to overcome the arbitrary of specifying a cluster mean or center which will greatly impact the accuracy of clustering, we propose a hierarchical clustering algorithm using blockmodeling (HCUBE), which employs the structural equivalence to measure the similarity of pages and reduces a large, potentially incoherent WSN to a smaller comprehensible structure based on the generated blocks of structural equivalence. HCUBE can find any clusters of arbitrary shapes in WSNs without a given hierarchical structure.
2. We use an Euclidean-like dissimilarity measure to compute the difference among pages in WSNs. And, we propose two new concepts called *inter-connectivity* and *closeness between clusters* and present how to compute these two measures. To improve the time performance of HCUBE, we introduce an optimized two-phase HCUBE algorithm.
3. We conduct experiments to estimate the accuracy and efficiency of HCUBE by comparing it with the *k*-means based clustering algorithm. Experimental results show that HCUBE is effective to partition WSNs and the optimized HCUBE algorithm can greatly reduce the time cost.

2. Related work

The problem of partitioning complex networks into clusters of objects has recently received increasing attention (Xu and Chen, 2005; Qin et al., 2005). In general, existing clustering methods can be classified into the following categories (Han and Kamber, 2000; Schaeffer, 2007):

1. Partitioning methods. The most popular heuristic methods are the *k*-means and the *k*-medoids algorithms. These two methods use the Euclidean distance between two objects to measure the clustering similarity, and work well for finding spherical-shaped clusters in small to medium sized data. To find clusters with complicated shapes or cluster very large data sets, partitioning methods need to be improved.
2. Hierarchical methods. A hierarchical method is categorized into the "bottom-up" and "top-down" approaches. These methods suffer from the fact that once the merge or split step is done, it can never be restored. A commonly used method is Chameleon, which is proposed by Karypis et al. (1999). It is a dynamic modeling in hierarchical clustering and its basic idea is that two clusters are merged if the inter-connectivity and proximity between two clusters are greatly near to the internal connectivity of objects within clusters. Chameleon is good at discovering arbitrary shaped clusters. However, it has the following drawbacks: (1) some important parameters should be specified manually, e.g., the *k* value in the *k*-nearest neighbor graph and the threshold of similarity measure function and (2) the processing cost for high dimensional data takes $O(n^2)$ time for n objects in the worst case. In order to overcome the first disadvantage, Long et al. (2009) proposed an improved Chameleon algorithm called M-Chameleon by applying the structural equivalence similarity degree and modularity theory to Chameleon algorithm. Du et al. (2007) partitioned the graph into initial clusters using information on vertex degrees and combined the initial clusters until an agreeable clustering is achieved. Donetti and Muñoz (2004) performed agglomerative clustering using spectral properties to construct the full cluster hierarchy and then selected a clustering from the resulting tree maximizing modularity. Flake et al. (2004) identified clusters by inserting an artificial sink and calculating flows to that sink. The minimum cuts that correspond to the maximum flows are used to build a minimum cut tree.
3. Density-based methods. In order to overcome the difficulty at discovering clusters of arbitrary shapes by distance-based partitioning methods, Ester et al. (1996) proposed a typical density-based approach called DBSCAN which grows clusters via a given density threshold.
4. Grid-based methods. A grid-based approach quantizes the object space into a finite number of cells, and then performs clustering operations on the grid structure. The main advantage of this method is the fast processing time that is independent of the number of objects. A typical example is STING (Wang et al., 1997) which is a grid-based multi-resolution approach.

To the best of our knowledge, there is rare cluster analysis approach suitable for WSNs. Actually, WSNs contain complex link relations that can be used to evaluate the similarity of pages, and we can group WSNs by analyzing the structure similarity among web pages. This motivates us to develop a new cluster approach that consider the structure relations among distinct pages. Blockmodeling is an alternative method to cluster WSNs due to the following two reasons: (1) it can partition pages of a WSN into a set of positions and the link relations into different blocks and (2) actors in one cluster have similar link relations with other parts of the network.

In terms of the studies in SNA, there are three key techniques including relational analysis, positional analysis, and hierarchical clustering (Xu and Chen, 2005). Unlike relational analysis, positional analysis examines how similarly two network actors connect to each other. The purpose of positional studies is to uncover the overall structure of a social network. An important method is blockmodeling (Batagelj et al., 2004), which aims to cluster objects that have substantially similar patterns of relations with others, and can interpret the patterns of relationships among clusters, especially fit for social networks. Although relational analysis and positional analysis approaches are based on different measures, they may employ hierarchical clustering to partition a social network. Hierarchical clustering generates a cluster hierarchy represented by a dendrogram, in which clusters are merged at successively less restrictive values of similarity (Xu and Chen, 2005).

The studies in blockmodeling focus mainly on theoretical foundations, there is rare work that has been done on employing hierarchical clustering to partition a network. Initially generalized blockmodeling approach is proposed by Batagelj, Doreian, and Ferligoj. This model is supported by two programs, that is, Model and Model2, but the procedures in these two models are time consuming and can only be applied to the networks of moderate size (Batagelj et al., 2004). Blockmodeling has recently attracted