Knowledge-Based Systems 000 (2017) 1–12

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Mining application-aware community organization with expanded feature subspaces from concerned attributes in social networks

Peng Wu[a,b], Li Pan[a,b,*]

[a] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Rd, Shanghai, China
[b] National Engineering Laboratory for Information Content Analysis Technology, Shanghai Jiao Tong University, Shanghai, China

ARTICLE INFO

Article history:
Received 28 March 2017
Revised 12 August 2017
Accepted 4 October 2017
Available online xxx

MSC:
68P10
91D30

Keywords:
Community detection
Semi-supervised clustering
Social networks

ABSTRACT

Social networks are typical attributed networks with node attributes. Traditional attribute community detection problems aim at obtaining the whole set of communities in the network. Different from them, we study an application-oriented problem of mining an application-aware community organization with respect to a set of specific concerned attributes. The set of concerned attributes is provided based on the requirements of any application by a user in advance. The application-aware community organization w.r.t. the set of concerned attributes consists of the communities whose attribute subspaces contain such set of concerned attributes. Besides concerned attributes, the subspace of each required community may contain some other relevant attributes. All relevant attributes of a subspace jointly describe and determine the community embedded in such subspace. Thus the problem includes two subproblems, i.e., how to expand the set of concerned attributes to complete subspaces and how to mine the communities embedded in the expanded subspaces. Two subproblems are jointly solved by optimizing a quality function called subspace fitness. An algorithm called ACM is proposed. In order to locate the communities potentially belonging to the application-aware community organization, a network backbone composed of nodes with similar concerned attributes is constructed. Then the cohesive parts of the network backbone are detected and set as the community seeds to locate the required communities. The set of concerned attributes is set as the initial subspace for all communities. Then each community and its attribute subspace are adjusted iteratively to optimize the subspace fitness. Extensive experiments on synthetic datasets demonstrate the effectiveness and efficiency of our method and applications on real-world networks show its application values.

1. Introduction

Community structure is one of the most prominent features of social networks, as it helps to visualize the network structures [1], enhance the information retrieval and promote the products recommendation [2], etc.. Social networks are typical attributed networks whose nodes are associated with attribute vectors. Traditional community detection methods [3–9] only consider structure cohesiveness property of communities. Recent attribute community detection methods [10–13] proposed for attributed networks require each detected community to be structurally dense and well separated from the rest of the network, as well as to have large similarity inside under a certain attribute subspace. In this paper, an attribute subspace consists of several attributes. Each subspace has a subspace vector that is used for attribute similarity computation. A subspace vector consists of importance weights of all attributes. The importance weights of those attributes that are not in a subspace are set as 0 in the corresponding subspace vector, because these attributes have no contribution to the attribute similarity under such subspace. A community has large similarity inside under a subspace when its nodes have similar values on the attributes in such subspace. Each community has a matched subspace containing all attributes on which nodes in such community have similar values, and the subspaces of different communities are usually different [10,11].

For a specific application, a set of communities with specific subspaces rather than all communities is usually required. In order to mine the set of communities required by a certain application, a user should provide some priori information in advance. It is hard to provide the complete subspaces of the required communities directly. In most cases, a user can only provide a handful of concerned attributes that the subspaces of the required communities should contain based on the requirements of specific applications. Besides the concerned attributes, the subspace of each required

* Corresponding author.
E-mail addresses: catking@sjtu.edu.cn, realcatking@163.com (P. Wu), panli@sjtu.edu.cn (L. Pan).

https://doi.org/10.1016/j.knosys.2017.10.007
0950-7051/

community may contain some other relevant attributes on which such required community also has similar values. In this paper, we study the problem of mining the set of communities whose subspaces contain the set of concerned attributes provided by a user for specific applications. Such set of communities is called application-aware community organization, or community organization for short. The subspaces of communities in the community organization are called feature subspaces. In our problem, the community organization mining is steered by the provided concerned attributes. Recently, several semi-supervised attribute community detection methods which can adjust the detection results based on user interests were proposed. FocusCO [14] aims to extract communities whose nodes are similar to the exemplar nodes that a user provides in advanced. DCM [15] mines a community with description from a provided community description which is defined as a query composed of disjunctions of conjunctions over basic conditions. The application scenarios of their problem are different from ours. The provided priori information of FocusCO is a set of exemplar nodes and that of DCM is a community description, while that of our problem is a set of concerned attributes. Moreover, our problem aims to mine the community organization whose communities have subspaces containing the concerned attributes, while FocusCO aims to mine communities whose nodes are similar to the exemplar nodes, and DCM aims to mine one community with the provided community description.

To solve the proposed problem, we put forward ACM, an Application-aware Community organization Mining method. A set of concerned attributes is provided by a user in advance. Since feature subspaces may contain some other implicit relevant attributes besides the provided ones, the set of provided concerned attributes should be expanded to complete feature subspaces. Each feature subspace and its embedded communities should match with each other. The subspace should make its embedded community have as large similarity inside as possible. Meanwhile, the community should be densely intra-connected and well separated from the rest of the network, as well as have as large similarity inside as possible under its subspace. The adjustment goals of each community and its subspace are similar. Thus, they are adjusted iteratively by optimizing a unified quality function called subspace fitness. The set of concerned attributes is set as the initial subspace. In order to locate the potential communities belonging to the community organization, edges between nodes having large similarity in initial subspace are sampled to construct a network backbone, and the cohesive parts of the network backbone are detected and set as the initial communities. Then each community and its subspace are adjusted based on each other to optimize the subspace fitness. Overlapping is an important characteristic for community structure in real-world networks [6–9]. The communities mined by our method can be naturally overlapping, because they are independently extracted by locally adjusting every initial community individually. However, highly overlapping communities often imply that they represent the same community [12]. Thus for highly overlapping communities, only one of them need to be reserved and others are redundant. The redundant communities are eliminated in our method.

The method proposed in this paper is flexible and applicable in many real-world application scenarios. For product recommendation in a co-purchased network, a user wants to recommend some products to a customer. He first figures out some concerned attributes of products that were bought by the customer before, and then provides these concerned attributes for ACM. The mined product community organization can be recommended to such customer. As another example, for product advertisement in a social network, a user wants to advertise a product in some customers. He first figures out some attributes of the potential customers, then these concerned attributes are provided for ACM and such product can be advertised in the mined customer community organization.

Our contributions are summarized as follows:

1. Considering a user can usually provide a handful of concerned attributes that the subspaces of the required communities should contain based on the requirements of specific applications, we propose a new problem of mining the set of communities whose subspaces contain the set of provided concerned attributes.
2. We define a unified quality function that can evaluate the quality of each community and its subspace simultaneously, and propose an algorithm ACM that locates the set of required communities and iteratively adjusts each community and its subspace to optimize the unified quality function.
3. We conduct extensive experiments on Synthetic and real-world datasets to evaluate the effectiveness and efficiency of ACM on proposed problem and show its application values.

The rest of this paper is organized as follows. Some related works are discussed in Section 2. Section 3 describes and models the proposed problem. Section 4 presents the greedy ACM algorithm in details. Experiments results are analyzed in Section 5. Finally, Section 6 concludes the paper.

2. Related work

Traditional community detection methods [3–9] which consider plain networks without attributes require each community to be structurally dense and well-separated from the rest of the network. Hajiabadi et al. [6] propose a generalized approach for community detection via a primary association node based criterion degree. Raj et al. [7] propose an overlapping community detection method by solving the stability-plasticity problem with a Fuzzy Adaptive resonance theory inspired algorithm. Wang et al. [8] use the PageRank algorithm to evaluate the node mass, and determine the community affiliation of nodes based on their positions in the inherent peak-valley structure of the topology potential field. Wang et al. [9] propose a Bayesian probabilistic model for automatic detection of communities in temporal networks and propose a gradient descent algorithm to optimize the objective function of their model. Traditional community detection methods which only consider structure cohesiveness property of communities may fail to detect communities with large attribute similarity inside in attributed networks.

Most of attribute community detection methods for attributed networks take unsupervised clustering techniques. In the early stage of the development, they treat all available attributes as equally important and consider the attribute similarity of each community under attribute full space [16–24]. The SA-Cluster [16,18] and its extended version Inc-Cluster [17,19] define a unified neighborhood random walk distance on an augmented graph by considering all available attributes as additional attribute vertices, and then take a K-Medoids method to cluster the network based on this unified distance. CESNA method [22] and BAGC method [21,25] statistically model the link structure and all available node attributes, and then obtain communities by inferring parameters of their statistical models. PICS method [20] defines an encoding cost used to describe the adjacency matrix and attribute matrix of a network and gets communities by minimizing the cost. CODICIL method [23] creates content edges based on content similarity and combines content edges with structure edges. Then it samples edges that are locally relevant for each node and clusters the resulting backbone network by any standard community detection method. With the increasing dimensionality of attribute space, the discrimination power of the attribute distance or similarity in full space may decrease [12]. Thus attribute subspace community

Please cite this article as: P. Wu, L. Pan, Mining application-aware community organization with expanded feature subspaces from concerned attributes in social networks, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.10.007