ARTICLE IN PRESS

Information Sciences xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Description-oriented community detection using exhaustive subgroup discovery

Martin Atzmueller *, Stephan Doerfel, Folke Mitzlaff

University of Kassel, Research Center for Information System Design, Knowledge and Data Engineering Group, Wilhelmshöher Allee 73, 34121 Kassel, Germany

ARTICLE INFO

Article history:
Received 10 April 2014
Received in revised form 25 April 2015
Accepted 3 May 2015
Available online xxxx

Keywords:
Social network analysis
Community detection
Subgroup discovery
Optimistic estimates
Exceptional model mining
Exploratory data analysis

ABSTRACT

Communities can intuitively be defined as subsets of nodes of a graph with a dense structure in the corresponding subgraph. However, for mining such communities usually only structural aspects are taken into account. Typically, no concise nor easily interpretable community description is provided.

For tackling this issue, this paper focuses on description-oriented community detection using subgroup discovery. In order to provide both structurally valid and interpretable communities we utilize the graph structure as well as additional descriptive features of the graph's nodes. A descriptive *community pattern* built upon these features then describes and identifies a community, i.e., a set of nodes, and vice versa. Essentially, we mine patterns in the "description space" characterizing interesting sets of nodes (i.e., subgroups) in the "graph space"; the interestingness of a community is evaluated by a selectable quality measure.

We aim at identifying communities according to standard community quality measures, while providing characteristic descriptions of these communities at the same time. For this task, we propose several optimistic estimates of standard community quality functions to be used for efficient pruning of the search space in an exhaustive branch-and-bound algorithm. We demonstrate our approach in an evaluation using five real-world data sets, obtained from three different social media applications.

1. Introduction

While classic community detection, e.g., [17] for a survey, just identifies subgroups of nodes with a dense structure, lacking an interpretable description, this paper focuses on the task of *description-oriented community detection*. Using additional descriptive features of the nodes contained in the network, we approach the task of identifying communities as sets of nodes together with a *description*, i.e., a logical formula on the values of the nodes' descriptive features. Such a *community pattern* then provides an intuitive description of the community, e.g., by an easily interpretable conjunction of attribute–value pairs. This is usually not achieved by classical community mining methods that consider the nodes of a network (e.g., denoting users in a social network) as mere strings or ids.

We present an algorithm for description-oriented community detection of the top-k communities (described by community patterns) with respect to a number of standard community evaluation functions. The method is based on an adapted

* Corresponding author.
E-mail addresses: atzmueller@cs.uni-kassel.de (M. Atzmueller), doerfel@cs.uni-kassel.de (S. Doerfel), mitzlaff@cs.uni-kassel.de (F. Mitzlaff).

http://dx.doi.org/10.1016/j.ins.2015.05.008
0020-0255/© 2015 Elsevier Inc. All rights reserved.

subgroup discovery approach [10,36], and also tackles typical problems that are not addressed by standard approaches for community detection such as pathological cases like small community sizes. We focus on interpretable patterns that can easily be incorporated into a practical application, for example, for recommendations in social bookmarking systems. It is important to note that we focus on static social graphs and do not take the dynamics into account since we aim to characterize a given community (allocation) for a given fixed interaction structure. Also, since in practice the entities in a network tend to belong to a number of different communities, the presented method naturally captures overlapping community allocations. Moreover, in contrast to global approaches, we focus on the discovery of local communities. According to the idea of local pattern mining, e.g., [20], we do not try to find a complete (global) partitioning of the network. Instead, we consider a set of local, potentially overlapping communities. These should be as exceptional as possible with respect to a given community quality measure.

We demonstrate our approach on several social media applications such as social networking and social bookmarking systems that provide interaction networks like explicit friendship relations between users. However, the presented approach is not limited to such systems and can be applied to any kind of graph-structured data for which additional descriptive features (node labels) are available, e.g., certain activity in telephone networks or interactions in face-to-face contacts [6] that also utilize tags or topic descriptions for the contained relations.

As an accompanying example, throughout the paper we use the friendship graph of the social bookmarking system BibSonomy[1] [15]. In BibSonomy, users can declare their friendship toward other users, thus, creating a directed graph with users as nodes. At the same time, each user collects and tags resources like publications and web pages. Thus, a user's set of tags can be considered as a description of that user's interests. The community mining task here is to find user groups, where users are well connected by their friendship links and share a common interest in one or more features (tags).

Overall, the contribution of this paper can be summarized as follows:

1. We first introduce description-oriented community detection and present the COMODO algorithm for obtaining the k-best community patterns using a given community evaluation measure. COMODO is a branch-and-bound algorithm based on an exhaustive subgroup discovery approach.
2. For fast description-oriented community detection using COMODO, we propose optimistic estimates [25,62] which are efficient to compute. We consider a number of standard community quality functions: The *segregation index* [19], the *inverse average ODF (out degree fraction)* [38], and the *modularity* [49]. We discuss the different measures for unweighted and weighted graphs, and extend the optimistic estimates accordingly.
3. We evaluate the presented approach using five data sets from three real-world social applications, i.e., from the social bookmarking systems BibSonomy and delicious,[2] and from the social media platform last.fm.[3]

The remainder of the paper is structured as follows: Section 2 summarizes basics of subgroup discovery, and provides general notions of graphs and community mining measures. Next, Section 3 introduces the proposed approach for description-oriented community detection and presents a number of optimistic estimates for standard community evaluation functions. After that, Section 4 discusses related work. For demonstrating the effectiveness and validity of the presented approach, Section 5 provides experiments using five data sets and discusses their results in the context of the three real-world applications. Finally, Section 6 concludes the paper with a summary and directions for future research.

2. Preliminaries

In the following, we briefly introduce basic notions with respect to pattern mining using subgroup discovery, graphs, and community quality measures.

2.1. Pattern mining using subgroup discovery

Subgroup discovery [28,62,13,5] aims at identifying interesting patterns with respect to a given target property of interest and according to a specific quality (interestingness) measure. The top patterns are then ranked according to the selected quality measure.

Formally, a database $D = (I, A)$ is given by a set of individuals I and a set of attributes A. A *selector* or *basic pattern* $sel_{a=a_j}$ is a boolean function $I \rightarrow \{0, 1\}$ that is true, iff the value of attribute a is equal to a_j for the respective individual. For a numeric attribute a_{num} selectors $sel_{a \in [min_j; max_j]}$ can be defined analogously for each interval $[min_j; max_j]$ in the domain of a_{num}. In this case, the corresponding boolean function is set to true, iff the value of attribute a_{num} is within the respective range. The set of all basic patterns is denoted by S.

[1] http://www.bibsonomy.org.

[2] http://www.delicious.com.

[3] http://last.fm.