Efficient Descriptive Community Mining

Martin Atzmueller and Folke Mitzlaff
Knowledge and Data Engineering Group,
University of Kassel, Germany
{atzmueller,mitzlaff}@cs.uni-kassel.de

Abstract

Community mining is applied in order to identify groups of users which share, e.g., common interests or expertise. This paper presents an approach for mining descriptive patterns in order to characterize communities in terms of their distinctive features: For an efficient discovery approach, we introduce optimistic estimates for obtaining an upper bound for the community quality. We present an evaluation using data from the real-world social bookmarking system BibSonomy.

Introduction

Community mining is a prominent approach for identifying densely connected subgroups of the nodes contained in a network. A community is intuitively defined as a set of nodes that has more and/or better links between its members compared to the rest of the network.

This paper proposes an approach for mining descriptive community patterns according to standard community evaluation measures: The proposed method collects patterns that describe communities by combinations of features, e.g., tags or topics for social bookmarking systems. We can consider, for example, groups of users interested in the topic *web mining*, *computer* and *java*. In this way, we aim to *identify and describe* interesting communities, in contrast to standard community mining approaches, e.g., (Newman 2004) that only identify communities as subsets of users. Our contribution is three-fold: We introduce the descriptive community mining scenario, and propose an approach for mining descriptive community patterns. Furthermore, we present optimistic estimates for pruning the search space, discuss their application in the context of standard community evaluation measures, and evaluate their impact.

Our application context is given by social and ubiquitous applications such as social networking applications, social bookmarking systems, and sensor-networks. Considering the BibSonomy system as an example, the friend graph indicates explicit friendship relations between users. Then, these graphs directly indicate communities (of users) according to the link structure. Similar interaction networks are obtained in the context of ubiquitous applications (e. g., users which are using a given service at the same place and time). Communities of users can then be characterized in terms of their descriptive features. In the context of social bookmarking systems, for example, we can consider the applied set of tags and the different resources (i.e., bookmarks, publications).

The rest of the paper is structured as follows: We first summarize basic notions of pattern mining, graphs, and according measures. Next, we discuss related work. After that, we introduce the approach for mining descriptive community patterns and describe optimistic estimates for standard community evaluation functions. Furthermore, we provide evaluation results of the presented approach in the context of the real-world BibSonomy system. Finally, we conclude the paper with a summary and directions for future research.

Preliminaries

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

Pattern Mining using Subgroup Discovery

Subgroup discovery (Wrobel 1997) aims at identifying interesting patterns with respect to a given target property of interest according to a specific quality function. Let Ω_A denote the set of all **attributes**. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined. Let DB be the **database** containing all available data records. A **data record** $r \in DB$ is given by the n-tuple $r = ((a_1 = v_1), \ldots, (a_n = v_n))$ of $n = |\Omega_A|$ attribute values, $v_i \in dom(a_i)$ for each a_i. A **subgroup description** $sd(s)$ of the subgroup s, $sd(s) = \{e_1, \ldots, e_l\}, l \geq 0$, is defined by the conjunction of a set of selection expressions (selectors). The individual selectors $e_i = (a_i, V_i)$ are selections on domains of attributes, $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. A **subgroup** s described by the subgroup description $sd(s)$ is defined as a subset of the whole database DB, i.e., $s \subseteq DB$: It is given by all records $r \in DB$ covered by the subgroup description $sd(s)$. We denote the subgroup s described by $sd(s)$ with $ext(sd(s))$. A subgroup s' is called a **refinement** of s, if $sd(s) \subset sd(s')$.

A **quality function** $q : 2^{DB} \rightarrow R$ assigns a numeric interestingness value to the subgroup s. For many quality functions an **optimistic estimate** of a subgroup s can be specified. This approximation describes an upper bound for the quality, that any refinement of s can have.

If the optimistic estimate of the current subgroup is below the quality of the worst subgroup of the k best subgroups obtained so far, then the current branch of the refinement tree can be safely pruned. More formally, an optimistic estimate oe of a quality function q is a function such that $s' \subseteq s \rightarrow oe(s) \geq q(s')$, i.e., that no refinement of subgroup s can exceed the quality $oe(s)$.

Graphs

A **graph** $G = (V, E)$ is an ordered pair, consisting of a finite set V which consists of the **vertices/nodes**, and a set E of **edges**, which are two element subsets of V. A *directed graph* is defined accordingly: E denotes a subset of $V \times V$. We write $(u, v) \in E$ in both cases for an edge belonging to E and freely use the term *network* as a synonym for a graph. The **degree** of a node in a network measures the number of connections it has to other nodes. For the **adjacency matrix** $A \in \mathbb{R}^{n \times n}$ of a set of nodes S with $n = |S|$ of a graph $G = (V, E)$ holds $A_{ij} = 1$ iff $(i, j) \in E$ for any nodes i, j in S (assuming some bijective mapping from $1, \ldots, n$ to S).

Community Quality Measures

The concept of a **community** can be intuitively defined as a group $\mathcal{C}$ of individuals out of a population $\mathcal{U}$ such that members of $\mathcal{C}$ are densely "related" one to each other but sparsely "related" to individuals in $\mathcal{U} \setminus \mathcal{C}$. This concept transfers to vertex sets $C \subseteq V$ in graphs $G = (V, E)$ where nodes in C are densely connected but sparsely connected to nodes in $V \setminus C$. For a given graph $G = (V, E)$ and a community $C \subseteq V$ we set $n := |V|$, $m := |E|$, $n_C := |C|$, $m_C := |\{(u, v) \in E \mid u, v \in C\}|$, $m_C := |\{(u, v) \in E \mid u \in C, v \notin C\}|$ and for a node $u \in V$ its degree is denoted by $d(u)$. Different evaluation functions $f \colon \mathcal{P}(V) \rightarrow \mathbb{R}$ for modeling the intuitive community concept exist, e. g., (Leskovec et al. 2008).

$$CON(C) = \frac{m_C}{2m_C + m_C} = 1 - \frac{2m_C}{\sum_{u \in C} d(u)} \quad (1)$$

In the context of this paper, we focus on maximizing local quality functions for single communities. The conductance CON compares the links *between* to the links *within* communities and is closer to zero for communities with higher quality: Therefore, in the following we consider the **inverse conductance** $COIN(C) = 1 - CON(C)$).

The modularity focuses on the number of edges *within* a community and compares that with the expected such number given a null-model (i.e., a randomized model). The modularity $MOD(S)$ of a set of nodes S and its assigned adjacency matrix $A \in \mathbb{N}^{n \times n}$ is given by

$$MOD(S) = \frac{1}{2m} \sum_{i,j} \big(A_{i,j} - \frac{d(i)d(j)}{2m}\big) \delta(C_i, C_j)\,, \quad (2)$$

where C_i is the cluster to which node i belongs and C_j is the cluster to which node j belongs; $d(i)$ and $d(j)$ denote i and j's degrees respectively; $\delta(C_i, C_j)$ is the *Kronecker delta* symbol that equals 1 iff $C_i = C_j$, and 0 otherwise.

The **local modularity** for a single community C can be computed as:

$$MODL(C) = \frac{1}{2m} \sum_{i \in C, j \in C} \big(A_{i,j} - \frac{d(i)d(j)}{2m}\big)\,.$$

Networks in Social Bookmarking Systems

In the following, we summarize three interaction networks that are provided by the BibSonomy system. All of these are typically also found in other resource sharing and social applications.

- The *Friend-Graph* $G_F = (V_F, E_F)$ is a directed graph with $(u, v) \in E_F$ iff user u has added user v as a friend.
- The *Click-Graph* $G_C = (V_C, E_C)$ is a directed graph with $(u, v) \in E_C$ iff user u has clicked on a link on the user page of user v.
- The *Visit-Graph* $G_V = (V_V, E_V)$ is a directed graph with $(u, v) \in E_V$ iff user u navigated to v's user page.

We refer to (Mitzlaff et al. 2010) for more details on the networks and a discussion concerning their application for community mining and assessment.

Related Work

Fortunato (Fortunato and Castellano 2007) discusses various aspects connected to the concept of community structure in graphs. A community detection method for a folksonomy is presented in (Kashoob, Caverlee, and Kamath 2010). Using a metric which is purely based on the structure of graphs, Newman presents algorithms for finding communities and assessing community structure in graphs (Newman 2004). (Adnan, Alhajj, and Rokne 2009) present an approach for community detection based on features identified by frequent pattern mining not considering the network structure.

In contrast to the approaches mentioned above, the proposed method integrates the information from both the network and other descriptive information, e.g., tags or topics describing the nodes contained in the network. The presented method focuses on the characterization and description of communities; it directly searches for the top k descriptive communities according to standard community evaluation measures. Therefore, the method is goal-directed by considering both the network (links) and the descriptive information for community mining.

Optimistic estimates for efficient knowledge discovery have been discussed, e.g., by (Wrobel 1997; Grosskreutz, Rüping, and Wrobel 2008) in the context of subgroup discovery. To the best of the authors' knowledge, no descriptive community mining approach applying such branch-and-bound methods has been proposed so far. A first approach for the characterization and description of communities was introduced in (Atzmueller et al. 2009), focussing on the description of spammers in the social bookmarking system BibSonomy. The proposed methods extends this using optimistic estimates for efficiently searching the description space while directly optimizing the community measures on the given network structure at the same time.

Mining Descriptive Community Patterns

In an intuitive sense, community mining is concerned with the identification of subgroups of users that are more densely connected internally than to other groups. Hence, subgroups and communities are rather similar, and we will use the terms interchangeably whenever we are referring to communities (e.g., users), either represented by a set of edges or nodes contained in a graph or dataset, respectively.

Overview

For the characterization of the communities, we consider a database DB containing records that describes a set of users, e.g., using topics the user is interested in, see Table 1 for some examples. Additionally, we consider links between the users modeled in a graph G, e.g., friendship links.

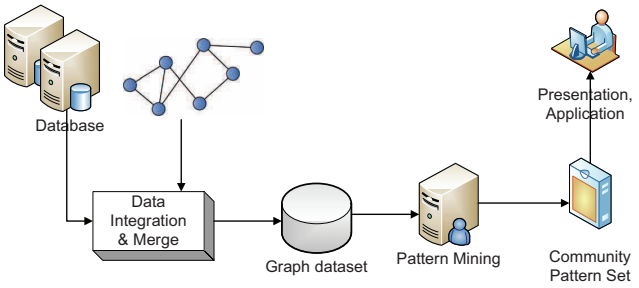

Figure 1: Overview on the presented aproach

For pattern mining we need a consolidated data representation. Therefore, we apply a data integration and merge step for obtaining a (flat) graph dataset, i.e., a dataset describing the nodes and edges in the graph. This dataset is constructed in a special way as outlined below. Using this new dataset, we apply the proposed pattern mining method utilizing optimistic estimates (of the standard quality functions) for pruning in order to provide an efficient approach. After the set of the k best community patterns has been obtained, it is ready for application, e.g., for presentation to the user for inspection, or for automatic approaches such as user recommendation or personalization of services. The approach is sketched in Figure 1. In the following, we discuss the data integration and pattern mining steps in detail.

Data Integration

Our goal is to discover the k best communities described by the attributes of the database DB, that maximize a community evaluation function with respect to G. Considering the DB and G, it is easy to see that both consider individual nodes, i.e., users. However, the community evaluation measures focus on the edges, i.e., the connections between the nodes in order to assess the community qualities

Therefore, we merge the two data sets into a single data set containing the connecting edges between the contained nodes that are constructed in a special way: Each data record represents a connecting edge between two nodes of the network. The attribute values of each such data record are then given by the intersection of the (non-default) attribute values of each node that is connected by the corresponding edge. For example, considering tags corresponding to binary attributes we only consider the *true* values of each attribute, e.g., indicating that a tag or a topic was applied by both users represented by the given nodes. The rationale behind using the intersection is based on the observation, that an edge (and its two nodes) can only contribute to a community described by a certain attribute value, if this respective attribute value is contained in the data records of the two nodes.

The edge data record also stores the contributing nodes and their respective degrees. Then, only using the number of edges contained in the community m_C, the total number of edges, and the respective node degrees $d(i)$ of the nodes $i \in C$ of the community, the local modularity for a community C can be directly computed as follows:

$$\begin{aligned} MODL(C) &= \frac{1}{2m}\sum_{i\in C, j\in C}\left(A_{i,j} - \frac{d(i)d(j)}{2m}\right) = \\ &= \frac{1}{2m}\sum_{i\in C, j\in C} A_{i,j} - \sum_{i\in C, j\in C}\frac{d(i)d(j)}{4m^2} = \\ &= \frac{1}{2m} 2m_C - \sum_{i\in C, j\in C}\frac{d(i)d(j)}{4m^2} = \\ &= \frac{m_C}{m} - \sum_{i\in C, j\in C}\frac{d(i)d(j)}{4m^2} \end{aligned}$$

Conductance can similarly be calculated using only the parameters mentioned above, cf., Equation 1.

For subgroup discovery, we are interested in maximizing the given quality function, which works well for the modularity while conductance is closer to zero for communities with higher quality. Therefore, from now on we consider the *inverse conductance* ($COIN$) instead of the conductance, for maximizing the quality values.

$$COIN(C) = 1 - CON(C) = \frac{2m_C}{\sum_{u\in C} d(u)}$$

Optimistic Estimates for Community Mining

In the following we introduce optimistic estimates for typical community evaluation functions, i.e., for the introduced inverse conductance and for the local modularity.

Modularity An optimistic estimate for the *local modularity* can be derived based on the number of edges m_C within the community:

$$oe(MODL(C)) = \begin{cases} 0.25, & \text{if } m_C \geq \frac{m}{2}, \\ \frac{m_C}{m} - \frac{m_C^2}{m^2}, & \text{otherwise.} \end{cases}$$

Proof We start with a reformulation of the modularity. An optimistic estimate can then be derived considering the number of edges m_C within the community. Also, note that $\sum_{i\in C} d(i) = 2m_C + m_C$, considering the degrees $d(i)$ of the nodes i contained in a community C.

Method(s)	Community description
Conductance	{work, flickr, delicious}, {university, bib, surabaya}, {php, web, internet}, {internet, all, emulation}
Modularity	{php, web, internet}, {innovation, business, forschung}
Conduct./Modularity	{work, flickr, delicious}

Table 1: Example for descriptive community patterns: Three of the top ranked subgroups/communities according to conductance and modularity together with their respective topic description, using the friend-graph data described in the evaluation below. The rows show the different communities, consisting of several topics as sets of tags in the rows of the table.

$$\begin{aligned} MODL(C) &= \frac{m_C}{m} - \sum_{i\in C, j\in C} \frac{d(i)d(j)}{4m^2} = \\ &= \frac{m_C}{m} - \frac{1}{4m^2}\sum_{i\in C} d(i) \sum_{j\in C} d(j) = \\ &= \frac{m_C}{m} - \frac{1}{4m^2}\sum_{i\in C} d(i)(2m_C + m_C) = \\ &= \frac{m_C}{m} - \frac{1}{4m^2}(2m_C + m_C)^2 \leq \\ &\leq \frac{m_C}{m} - \frac{1}{4m^2}(2m_C)^2 = \frac{m_C}{m} - \frac{m_C^2}{m^2} = \\ &= \hat{oe}(MODL(C)). \end{aligned}$$

Note that the optimistic estimate is only dependent on m_C, i.e., the number of edges covered by the community s. Therefore, every subgroup $s^* \subseteq s$ that is a refinement of s will cover at most m_C edges.

The function $\hat{oe}(MODL(C))$ is a concave function since its derivative function

$$\hat{oe}(MODL(C))' = \frac{1}{m} - \frac{2m_C}{m^2}$$

is monotonically decreasing. Therefore, the function has one maximum, at point $\frac{m}{2}$, for $m \neq 0$.

We consider two cases: If $m_C \geq \frac{m}{2}$, then the maximal modularity can be obtained at point $\frac{m}{2}$. Otherwise, for all $m_C < \frac{m}{2}$, $\hat{oe}(MODL(C))$ is decreasing in m_C, and thus $\hat{oe}(MODL(C))$ is an optimistic estimate for $MODL(C)$. This concludes the proof. □

Inverse Conductance For the *inverse conductance*, we need to consider a minimal support threshold $\mathcal{T}_n$ w.r.t. the community size (number of nodes) when computing the optimistic estimate:

$$oe(COIN(C)) = 1 - \frac{\sum_{i=1}^{\mathcal{T}_n} d(i)}{\sum_{u\in C} d(u)}$$

where $d(i)$ are the outgoing degrees of the nodes contained in the community C, sorted in ascending order, such that $d(i), i = 1 \ldots \mathcal{T}_n$ denotes the minimal $\mathcal{T}_n$ outgoing degrees of connected nodes contained in the community C.

Proof

$$\begin{aligned} COIN(C) &= \frac{2m_C}{\sum_{u\in C} d(u)} = \\ &= \frac{\sum_{u\in C} d(u) - m_C}{\sum_{u\in C} d(u)} = \\ &= 1 - \frac{m_C}{\sum_{u\in C} d(u)} \leq \\ &\leq 1 - \frac{\sum_{i=1}^{\mathcal{T}_n} d(i)}{\sum_{u\in C} d(u)} \\ &= oe(COIN(C)). \end{aligned}$$

As shown above, for a fixed m_C it follows that $oe(COIN(C)) \geq COIN(C)$. Since every subset $C' \subseteq C$ will cover at most m_C edges and the numerator of the last term ($\sum_{i=1}^{\mathcal{T}_n} d(i)$) is the minimum considering the outgoing edges for a minimal size of $\mathcal{T}_n$, $oe(COIN(C))$ is an optimistic estimate of $COIN(C)$. □

The optimistic estimate can be efficiently computed by traversing the set of nodes and collecting the outgoing node count for each node considering the endpoints of the edges.

Algorithmic Issues

For mining community patterns, we apply the COMODO algorithm. COMODO is an adaptation of the SD-Map* algorithm (Atzmueller and Lemmerich 2009) and applies a special data structure, i.e., an adapted frequent pattern tree (FP-tree), cf., (Han, Pei, and Yin 2000), containing extended information for efficiently mining community patterns. The FP-tree data structure can be regarded as a compressed data representation for the set of instances. COMODO utilizes the FP-tree structure (built in two scans of the database) to efficiently compute quality functions for all subgroups. If all the necessary information is compiled into the tree structure, then the community evaluation measures can be evaluated locally at each node: Thus, we essentially need to store all the appropriate information within the FP-nodes of the FP-Tree. The FP-tree contains the frequent FP-nodes in a header table, and links to all occurrences of the frequent selectors in the FP-tree structure. In this way, the parameters (of combinations) of selectors can be easily retrieved. Due to the limited space we refer to Han et al. (Han, Pei, and Yin 2000) for more details on FP-trees.

To efficiently compute the community evaluation functions together with their optimistic estimates for the community mining context COMODO stores additional information

in the FP-nodes of the FP-Tree, depending on the used quality function. Each FP-node of the FP-Tree captures information about aggregated edge information concerning the data base DB and the respective network. For each node, we store the following information:

- The selector corresponding to the attribute value of the FP-node. This selector describes the subgroup (given by a set of edges) covering the FP-node.
- The edge count m_C of the (partial) community represented by the FP-node, i.e., the aggregated count of all edges $E_C = \{(u, v) \in E : u, v \in C\}$ that are accounted for by the FP-node and its selector, respectively.
- The set of nodes $V_C = \{u : (u, v) \in E_C, u \in C, v \in C\}$ that are connected by the set of edges E_C of the FP-node.

The presented optimistic estimates enable efficient pruning strategies for determining upper bounds for the community evaluation measures. Applying these, COMODO can reorder, sort, and prune the current hypotheses during search for the top k patterns. During the traversal of the tree, and the refinement of the hypotheses (descriptions), all FP-nodes with an optimistic estimate below the minimal quality contained in the k best solutions so far can be pruned.

The result of the COMODO algorithm for mining descriptive community patterns is the set of the top k patterns according to the applied community evaluation function. These top k patterns directly correspond to different communities described by the respective patterns. Thus, the proposed (exhaustive) method guarantees that the top k communities are discovered, that can be represented using the given description space.

Evaluation

In the following, we first describe the data used for the evaluation. We used publicly available data from the social bookmark and resource sharing system BibSonomy. After that, we present the conducted experiments and discuss the experimental results.

Evaluation Data and Setting

Our primary resource is an anonymized dump of all public bookmark and publication posts until January 27, 2010, from which we extracted *explicit* and *implicit* relations, cf. Table 2 for an overview.

The dump consists of 175,521 tags, 5,579 users, 467,291 resources and 2,120,322 tag assignments. The BibSonomy dump also contains friendship relations modeled in BibSonomy concerning 700 users. Furthermore, we obtained data extracted from the "*click log*" of BibSonomy, consisting of entries which are generated whenever a logged-in user clicked on a link in BibSonomy, and the Apache log entries.

Before performing the experiments, we applied *latent dirichlet allocation* (LDA) (Blei, Ng, and Jordan 2003) for data preprocessing, since using conjunctive community descriptions is very difficult using the whole set of tags since the respective data is rather sparse. Furthermore, there are several issues when utilizing the (raw) set of tags directly, e.g., relating to many synonyms, writing variations, and hierarchical dependencies between tags that need to be handled appropriately in order to get more meaningful results.

	G_V (Visit)	G_C (Click)	G_F (Friend)
$\lvert V_i \rvert$	3381	1151	700
$\lvert E_i \rvert$	8214	1718	1012
$\lvert V_i \rvert / \lvert U \rvert$	0.58	0.20	0.12

Table 2: High level statistics for all relations where U denotes the set of all users in BibSonomy.

LDA builds topics, as interpretable tag clusters, i.e., for obtaining descriptive topic consisting of associated sets of tags. A user u is thus represented as a vector $\vec{u} \in \mathbb{R}^{T'}$ in the topic vector space, where $T' \ll T$ is the number of topics. We applied datasets containing 100 (LDA-100) and 500 (LDA-500) topics each for the user – tag/topic relations.

Results and Discussion

During our experiments, we could directly observe the pruning potential provided by the proposed optimistic estimates. The significant reduction of the search space using the optimistic estimate functions is shown in Table 3. The table shows the reduction concerning the steps/hypotheses during the mining process using the optimistic estimates for local modularity and conductance. For the LDA-100 dataset the unpruned search space contains about $7.74 \cdot 10^8$ steps for the friend graph, and about $7.94 \cdot 10^8$ steps for the visit and click graphs. For the LDA-500 dataset, the search space contains about $3.6 \cdot 10^8$ steps for the friend graph and about $2.45 \cdot 10^{10}$ steps for the click and visit graphs. It is easy to see, that the optimistic estimates enable an efficient and tractable mining approach with significant pruning options.

While the optimistic estimate for local modularity enables a significant pruning at a very low minimal support level, i.e., for about 1% (support count $\mathcal{T}_n = 5$), the optimistic estimate for (inverse) conductance enables the pruning at higher support levels, e.g., $\mathcal{T}_n \geq 10$ similar to a 2% support level.

This can be explained by the fact that the local modularity (and its optimistic estimate) gives more importance to the size of the community (relating to the number of edges that are contained in the community), while the conductance only considers the fraction of the edges in and the edges leaving the community. In this way, very small communities can also obtain a high quality value according to the conductance, if the minimal support threshold is reached. This is especially relevant for LDA-100 and for denser graphs, e.g., the visit graph, cf. Table 2 for its characteristics. We also investigated the impact of the minimal support threshold compared to the combination with optimistic estimate pruning

	G_F		G_C		G_V	
	L100	L500	L100	L500	L100	L500
MODL	76.4	71.4	83.8	74.8	82.3	72.5
COIN	75.1	70.0	81.8	73.7	82.6	72.6

Table 4: Mean cosine-similarities (in percent) of the top 25 communities discovered by COMODO ($\mathcal{T}_n = 20$) given by the means of the respective pairwise node similarities.

Reduction in steps: Pruning with optimistic estimates of local modularity and iconductance

	G_F			G_C			G_V		
	MODL	*COIN*		*MODL*	*COIN*		*MODL*	*COIN*	
method	$\mathcal{T}_n = 5$	$\mathcal{T}_n = 10$	$\mathcal{T}_n = 20$	$\mathcal{T}_n = 5$	$\mathcal{T}_n = 20$	$\mathcal{T}_n = 50$	$\mathcal{T}_n = 5$	$\mathcal{T}_n = 50$	$\mathcal{T}_n = 70$
L100-K25	99.99	37.60	99.25	99.99	93.80	99.99	99.99	73.13	99.80
L100-K50	99.96	36.80	99.17	99.98	92.70	99.99	99.98	72.65	99.76
L100-K100	99.58	35.98	98.20	99.97	90.95	99.99	99.94	72.60	99.70
L500-K25	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99	99.99
L500-K50	99.98	99.98	99.99	99.99	99.98	99.99	99.99	99.99	99.99
L500-K100	99.98	99.97	99.99	99.99	99.97	99.99	99.99	99.99	99.99

Table 3: Impact of optimistic estimate pruning for local modularity and iconductance (Pruned steps in percent). L*x*-K*y* denotes the LDA-*x* dataset and the *y* top communities, for different support counts ($\mathcal{T}_n$). 99.99% of the steps are pruned, for example, using the optimistic estimate for local modularity for the LDA-100 dataset concerning the 25 top communities with $\mathcal{T}_n = 5$.

for the conductance quality function in more detail: As expected, the results indicate that the conductance optimistic estimate pruning enable a significant gain in pruning performance, after pruning with minimal support thresholds, even for larger ones. Considering the friend graph, for example, for a minimal support threshold $\mathcal{T}_n = 10$ we could observe a boost from 9.60% to 37.60%, and from from 87.80% to 93.80%, for the click graph with $\mathcal{T}_n = 20$, respectively.

Furthermore, we obtained the mean pairwise cosine-similarities, e.g., (Salton 1989), of the nodes contained in the communities for the different networks, and measures, cf., Table 4, according to the set of assigned topics. As expected, the similarity values for the LDA-100 dataset are higher than the ones for the (sparser) LDA-500 dataset. The results indicate a high similarity for users in the discovered communities, evidencing good community structure.

Conclusions

In this paper, we have presented an approach for mining descriptive community patterns: We discussed the descriptive setting and described how to efficiently perform the mining using optimistic estimates for standard community evaluation functions. The presented approach was evaluated using data from the social bookmarking system BibSonomy.

For future work, we aim to apply the proposed method on more (diverse) evidence networks. Additionally, we aim to analyze and compare further community quality functions regarding their impact and pruning potential.

Acknowledgements

This work has been partially supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University, and by the Commune project funded by the Hertie foundation. We thank the reviewers for their comments, and especially wish to thank Stephan Doerfel for helpful discussions.

References

Adnan, M.; Alhajj, R.; and Rokne, J. 2009. Identifying Social Communities by Frequent Pattern Mining. In *Proc. 13th Intl. Conf. Information Visualisation*, 413–418. Washington, DC, USA: IEEE Computer Society.

Atzmueller, M., and Lemmerich, F. 2009. Fast Subgroup Discovery for Continuous Target Concepts. In *Proc. 18th International Symposium on Methodologies for Intelligent Systems (ISMIS 2009)*. Springer Verlag.

Atzmueller, M.; Lemmerich, F.; Krause, B.; and Hotho, A. 2009. Who are the Spammers? Understandable Local Patterns for Concept Description. In *Proc. 7th Conference on Computer Methods and Systems*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

Fortunato, S., and Castellano, C. 2007. *Encyclopedia of Complexity and System Science*. Springer. chapter Community Structure in Graphs.

Grosskreutz, H.; Rüping, S.; and Wrobel, S. 2008. Tight optimistic estimates for fast subgroup discovery. In *Proc. ECML/PKDD 2008*, 440–456. Berlin: Springer Verlag.

Han, J.; Pei, J.; and Yin, Y. 2000. Mining Frequent Patterns Without Candidate Generation. In *2000 ACM SIGMOD Intl. Conference on Management of Data*, 1–12. ACM Press.

Kashoob, S.; Caverlee, J.; and Kamath, K. 2010. Community-Based Ranking of the Social Web. In *Proc21st ACM Conference on Hypertext and Hypermedia*.

Leskovec, J.; Lang, K. J.; Dasgupta, A.; and Mahoney, M. W. 2008. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters. http://arxiv.org/abs/0810.1355.

Mitzlaff, F.; Atzmueller, M.; Benz, D.; Hotho, A.; and Stumme, G. 2010. Community Assessment using Evidence Networks. In *Proc. Workshop on Mining Ubiquitous and Social Environments (MUSE2010)*.

Newman, M. E. J. 2004. Detecting Community Structure in Networks. *Europ Physical J* 38.

Salton, G. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley.

Wrobel, S. 1997. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, 78–87. Berlin: Springer Verlag.