Towards Understanding Spammers – Discovering Local Patterns for Concept Description

Martin Atzmueller, Florian Lemmerich, Beate Krause, and Andreas Hotho

University of Würzburg,
Department of Computer Science VI
Am Hubland, 97074 Würzburg, Germany
{atzmueller, lemmerich, krause, hotho}@informatik.uni-wuerzburg.de

Abstract. Concept description is an important task of descriptive data mining: Basically, its aim is to identify and to summarize properties of a selected target population in the form of a set of patterns – in a concise and comprehensible way. In this paper we present an approach for concept description in the social bookmarking domain: We show how subgroup discovery can be utilized for identifying discriminative and characteristic local patterns in order to understand the behavior of (non-)spammers. A case study applying data from a real-world system for social bookmarking provides exemplary results and demonstrates the applicability and effectiveness of the presented approach.

1 Introduction

Concept description or class description, e.g., [1], is an important method used in descriptive data mining that comprises two subtasks: *Concept characterization* aims to summarize a given target population in terms of typical or characteristic features. In contrast, *concept/class discrimination* generates descriptions comparing the target population to one or more contrasting populations. In this way, both techniques aim to describe the target population in complementing ways: Concept discrimination focuses on the differences between classes, by contrasting their *discriminating* features. On the other hand, concept characterization focuses on the common or *typical* features of a certain class. As we will see later, there is a trade-off between the characterization and discrimination goals, when considering the respective patterns.

Subgroup discovery [2, 3], is a broadly applicable technique for identifying properties of a selected target population: It is usually applied for data exploration and descriptive induction, in order to identify relations between a dependent (target) variable and usually many independent variables, e.g., consider the subgroups "users which take a longer time till their first post and have no middle name indicate spammers [target variable]" or accordingly "users with a low number of tags and an IP in range X are usually non-spammers". Due to its flexible discovery strategy applying an arbitrary and user-definable quality function, subgroup discovery is easily adaptable for discovering local patterns in the context outlined above. While subgroup discovery is commonly applied for concept/class discrimination, by identifying discriminating descriptions of subgroups with a (significantly) deviating distribution of the target (class), we will show how to extend and adapt the method for concept characterization.

In contrast to existing approaches for concept description (e.g., [4, Ch. 4.3]), subgroup discovery provides the following distinctive features: It can cope with a large number of relevant attributes and its search strategy is goal-oriented applying an arbitrary quality function that can be flexibly defined by the user.

Since subgroup discovery methods are not necessarily covering approaches, several of the discovered patterns can cover the same set of instances. In order to present a selected discriminative high-quality set of subgroups, the patterns should in general have an individual high quality, and a low overlap with respect to other competing patterns. Thus, the result set of the discovered subgroups can often be reduced by removing irrelevant subgroups [5, 6].

In this paper, we present an approach for concept characterization and discrimination using local patterns that are obtained using subgroup discovery techniques. We describe the general approach, and also discuss the relation of concept description and information retrieval, since in many applications (such as in the case study of this paper) the prominent evaluation metrics come from this context. Since a characterization task aims to cover the target group as comprehensibly as possible (recall) and a discrimination provides distinctive features of the target population (precision), both requirements can be captured by the applied method.

The context of the proposed approach is the description of (non-)spammers, i.e., characterizing spammers and non-spammers by using their distinctive features but also their common features (with respect to the respective (non-)spammer class). We show how local patterns for concept description can be retrieved, and we present a case study applying real-world data from the social bookmarking system BibSonomy [7].

The rest of the paper is structured as follows: Section 2 provides the necessary background, introducing subgroup discovery, and the method for removing irrelevant local patterns. Next, Section 3 describes the approach for discovering local patterns for concept characterization and discrimination in detail. After that, Section 4 presents a case study of the proposed approach in the BibSonomy domain. Section 5 discusses related work. Finally, Section 6 concludes the paper with a summary and interesting directions for future research.

2 Preliminaries

In the following, we first introduce the necessary notions concerning the used knowledge representation, before we define the subgroup discovery setting, and describe a method for removing irrelevant patterns.

2.1 Basic Definitions

Let Ω_A be the set of all attributes. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined. Furthermore, we assume $\mathcal{V}_A$ to be the (universal) set of attribute values of the form $(a = v)$, where $a \in \Omega_A$ is an attribute and $v \in dom(a)$ is an assignable value. We consider nominal attributes only so that numeric attributes need to be discretized accordingly.

Let CB be the case base (data set) containing all available cases, also often called instances. A case $c \in CB$ is given by the n-tuple $c = ((a_1 = v_1), (a_2 = v_2), \ldots, (a_n = v_n))$ of $n = |\Omega_A|$ attribute values, $v_i \in dom(a_i)$ for each a_i.

2.2 Subgroup Discovery Basics

The main application areas of subgroup discovery are exploration and descriptive induction, to obtain an overview of the relations between a (dependent) target variable and a set of explaining (independent) variables. Then, the goal is to uncover properties of the selected target population of individuals featuring the given target property of interest. Therefore, not necessarily complete relations but also partial relations, i.e., (small) subgroups with "interesting" characteristics can be sufficient.

A subgroup discovery task mainly relies on the following four main properties: the target variable, the subgroup description language, the quality function, and the search strategy. An efficient search strategy is necessary since the search space is exponential concerning all the possible selectors of a subgroup description. Often, heuristic beam search methods but also efficient exhaustive algorithms, e.g., [8], are applied.

In this paper we focus on binary target variables. For multiple classes/concepts we can equivalently generate multiple binary class-problems. Similar to the *MIDOS* approach [3] we try to identify subgroups that are, e.g., as large as possible, and have the most unusual (distributional) characteristics with respect to a given concept of interest represented by the target variable. The description language specifies the individuals belonging to the subgroup. The subgroup is thus given by all cases in the data set that satisfy its subgroup description. For a commonly applied single-relational propositional language a subgroup description can be defined as follows:

Definition 1 (Subgroup Description). *A subgroup description $sd = \{e_1, e_2, \ldots, e_n\}$ 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, $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. We define Ω_E as the set of all selection expressions and Ω_{sd} as the set of all possible subgroup descriptions.*

A quality function measures the interestingness of the subgroup and is used to rank these. Typical quality criteria include the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size.

Definition 2 (Quality Function). *Given a particular target variable $t \in \Omega_E$, a quality function $q : \Omega_{sd} \times \Omega_E \to R$ is used in order to evaluate a subgroup description $sd \in \Omega_{sd}$, and to rank the discovered subgroups during search.*

In comparison to the strict support/confidence framework applied for association rule mining, e.g., [9], the subgroup quality functions can be flexibly defined which provides for a powerful tool. Typical quality functions, cf., [2], include the deviation of the target share in the subgroup compared to the general population. For binary target variables, examples for quality functions are given by

$$q_{BT} = \frac{(p - p_0) \cdot \sqrt{n}}{\sqrt{p_0 \cdot (1 - p_0)}} \cdot \sqrt{\frac{N}{N - n}}\,, \quad q_{RG} = \frac{p - p_0}{p_0 \cdot (1 - p_0)}\,,$$

where p is the relative frequency of the target variable in the subgroup, p_0 is the relative frequency of the target variable in the total population, $N = |CB|$ is the size of the total population, and n denotes the size of the subgroup.

In contrast to the quality function q_{BT} (the classic binomial test), the quality function q_{RG} only compares the target shares of the subgroup and the total population measuring the *relative gain*. Therefore, a support threshold $\mathcal{T}_{Supp}$ is necessary to discover significant subgroups.

In addition to the parameters discussed above, other parameters, like the simplicity of the patterns can be flexibly included into the quality function. Below we will describe formalizations and adaptations of quality functions for concept description.

2.3 Removing Irrelevant Patterns

The result of subgroup discovery is a set of subgroups. Since subgroups can overlap, relevancy analysis is essential in order to identify a compact but comprehensive set of subgroups. As proposed in [10], we consider a subgroup s as more relevant with respect to another subgroup description s', if it covers all positive examples of s', but no negative examples, that are not covered by s' as well. Formally s' is irrelevant with respect to s, if and only if

$$TP(s') \subseteq TP(s) \text{ and } FP(s) \subseteq FP(s'),$$

where $TP(s)$ is the set of positive examples (containing the target concept) in the subgroup s and $FP(s) = s \setminus TP(s)$ denotes the set of negative examples contained in s.

To identify relevant subgroups we utilize efficient methods for mining relevant patterns as described in [5], based on an efficient vertical bitset-driven subgroup discovery algorithm, and the definition of relevant/irrelevant patterns for class-labeled data [10, 11]. The applied approach allows for the efficient and effective filtering of irrelevant patterns during the subgroup discovery process in order to obtain a more diverse set of subgroups and can also be adapted to incorporate the handling of exceptions into the strict relevancy definition.

3 Local Patterns for Concept Characterization and Discrimination

Concept description is one of the major tasks of descriptive data mining: Also often referred to as class description, it aims to describe a set of individuals in a concise and compact way, cf., [1]. In contrast to common descriptive data mining tasks such as association rule mining, e.g., [9], concept description and subgroup discovery can both be regarded as supervised learning tasks since they focus on a specific concept/class of interest, or generally on a certain property that we are interested in.

The concept description task can be focused on *characterization*, that is, discovering typical properties of the target concept, or on *discrimination*, that is, identifying the properties discriminating between the target concept and the non-target elements. We discuss these subtasks in detail below, and we show, how both can be mapped to a subgroup discovery task using suitable quality functions. Before that, we define *pattern rules* as a convenient representation formalism for subgroups and subgroup patterns, respectively.

3.1 Pattern Rules

Subgroup patterns, i.e., subgroups with their associated subgroup descriptions can be represented similar to rules containing the subgroup description in the body of the rule, and the target concept in the head of the rule. Since a subgroup is ranked by a quality function with respect to a specific target property, it can directly be assigned a quality rating, e.g., according to the value of the quality function. In the following, we define *pattern rules*, i.e., rules for formalizing subgroup patterns in a rule-like manner and an associated quantitative quality rating.

Definition 3 (Pattern Rule). *A pattern rule* $r = B(r) \rightarrow H(r)\ [q(r)]$ *is defined by the body* $B(r)$ *and the head* $H(r)$ *of the pattern rule, where* $B(r) \subseteq \Omega_E, H(r) \subseteq \Omega_E$ *for which the selectors are combined conjunctively, i.e.,* $e_1 \wedge \cdots \wedge e_k$*, (*$e_i \in \Omega_E, i = 1 \ldots k$*). A quality parameter* $q(r) \in \mathbb{R}$ *is assigned to the pattern* r *denoting its respective quality.*

3.2 Concept Characterization using Subgroup Patterns

The goal of concept characterization is to provide a concise and succinct summary of a given target concept: By identifying characteristics of a selected target population typical properties of the concept can be identified. Thus, the focus is on typical, or *necessary* features that occur for (almost) all objects of the concept. In that context, a necessary feature $f \in \Omega_{sd}$ means, that if target concept $t \in \Omega_E$ is observed, then f is also observed, i.e., $t \rightarrow f$. So, essentially necessary features are contained in all cases of the target concept. However, this usually only happens for few interesting features. Therefore, we will relax this condition for *near-necessary* features: We apply a quality measure q_c for expressing the typicality of each pattern with respect to the target concept. Then, we can propose a *characteristic* pattern rule $r = t \rightarrow f\ [q_c(r)]$ for the pattern that specifies the typicality of the pattern using the quality measure q_c.

3.3 Concept Discrimination using Subgroup Patterns

In contrast to concept characterization the aim of concept discrimination is to compare or to contrast a given target concept with one or more discriminating concepts. In order to obtain a comparative summary of the concept the distributions of the target and the contrasting concepts are compared in the different subpopulations. A perfect contrast is then given by *sufficient* properties or features: If feature $f \in \Omega_{sd}$ is observed then the target class $t \in \Omega_E$ is also observed, i.e., $f \rightarrow t$. So, such a rule also provides for certain classification given the feature. However, we need to relax this condition using *near-sufficient* features: Similar to concept characterization the discriminating patterns are also assigned a quality parameter based on their discriminative power, which can be used for formalizing *discriminative* pattern rules.

It is easy to see that subgroup discovery can be naturally applied for concept discrimination, since we consider pattern of the type $Body \rightarrow T$ in that case: We can simply map a given subgroup description $sd = \{e_1, e_2, \ldots, e_n\}$ to a *discriminative* pattern rule $r = e_1 \wedge e_2 \cdots \wedge e_n \rightarrow t\ [q(r)]$, with respect to the target concept t. Then, the quality parameter $q(r)$ is determined by the applied quality function q_d.

3.4 Quality Functions for Concept Description

In the following we describe the relevant parameters of a subgroup pattern. Furthermore, we discuss how the different quality functions can be applied both for concept characterization and concept discrimination. In a discriminative setting, let us consider a subgroup s, and its equivalent pattern rule $r = sd \rightarrow t\ [q(r)]$ with the subgroup description sd of s, and the target variable t. We construct two binary variables T and SD for the target class cases, and the cases covered by the subgroup description, respectively. We can then create a four-fold contingency-table as shown below.

	$T = true$	$T = false$
$SD = true$	tp	fp
$SD = false$	fn	tn

Considering the possible outcomes of the rules, we distinguish the *true positives (tp)* for which the pattern correctly predicts the target variable, the *false positives (fp)* for which the prediction is incorrect, and equivalently the *false negatives (fn)* and *true negatives* for the 'negation' of the rule, i.e., for the complement of the prediction. For a pattern rule $r' = t \rightarrow sd\ [q_c(r')]$, i.e., a rule that contains the target concept in the body of the rule, the entries for *fp* and *fn* are simply swapped in the contingency table.

So, the subgroup discovery approach can in principle be directly applied to both settings, since we just need to insert the right parameters into the quality functions: For the discriminative setting the target share p and the subgroup size n can be easily obtained using the parameters contained in the contingency table, since $p = tp/(tp + fp)$ and $n = tp + fp$. Note that this is equivalent to the *precision* of the pattern known from information retrieval, e.g., [12]. Furthermore, for concept characterization, we obtain an adapted $p' = tp/(tp + fn)$ and $n' = tp + fn$ since the 'reference subgroup' consists of all the positives of the total population. This is equivalent to the *recall* of the pattern known from information retrieval, if we consider the subgroup description as the 'query' and the instances covered by the subgroup as the result set.

For the discriminative setting we can readily apply all of the usual quality functions used for subgroup discovery. We can consider the relative gain quality function q_{RG}, for example, which is order-equivalent to *precision* discussed above. Then, we estimate the deviation of the distribution of the target concept in the subgroup compared to the general population.

For the characteristic setting, we need to apply special quality functions since we mainly want to characterize a selected target population, i.e., all the positives of the target concept. Therefore, we focus on the *positive coverage*, i.e., on the coverage of the target space, in contrast to discriminative quality functions that take the size of the subgroup into account corresponding to the coverage of the total population. Thus, for the characterization task, the distribution of the true positives tp needs to be compared to the total positives which can be obtained as $Pos = tp + fn$. For this purpose, for example, the quality function q_{TPR} measuring the true positive rate (or equivalently the recall) can be applied,

$$q_{TPR} = \frac{tp}{Pos} = \frac{tp}{tp + fn}\,,$$

which compares the true positives of the subgroup to all positives with respect to the target variable. This quality function is equivalent to the function proposed by Han [4, Ch. 4.3], and can be used for estimating the typicality of the subgroup with respect to the target population. Then, subgroups with a large overlap with the target class instances are selected, without considering the (potentially large) overlap with the non-target instances.

Concept description also concerns the understandability or comprehensibility of the patterns. Therefore, also the simplicity of the (rule) patterns is a major concern, cf., [13, 14]; the syntactic simplicity can be estimated using the length of the descriptions of the patterns. In order to incorporate this parameter into the applied quality measures $q*$, we simple take the fraction of the original quality function value $q(s)$ of the subgroup s and the length $|sd(s)|$ of the subgroup description $sd(s)$:

$$q^*(s) = \frac{q(s)}{|sd(s)|} \, .$$

In this way, patterns that are described by shorter (and thus simpler) descriptions will be favored. It is easy to see, that this is especially useful in the case of breaking ties between sets of subgroups with an equal quality ranking.

3.5 Concept Description Views

As discussed above, we can directly apply subgroup discovery for both subtasks of concept description, that is, concept characterization and discrimination by choosing suitable quality functions. These provide different *views* on the concept description task, because the analyst can focus on the *characterization* and on the *discrimination* aspect. However, in order to take both into account, we need to consider both the characterization and the discrimination, because for description we want to obtain 1) a large coverage of the *target concept*, and 2) a significant deviation of the target distribution within the subgroup pattern and the whole data set. Essentially, the analyst can navigate between the characteristic and discriminative patterns easily, e.g., by tweaking the threshold parameters, by modifying the set of analyzed attributes, or by adapting the applied quality function.

However, for a quick comprehensive view there is also an integrated option: As discussed above, the quality functions for characterization are similar to measuring the *recall* of the pattern, while the discriminative setting provides pattern with a high *precision*. Analogously to information retrieval, we can therefore combine quality functions for characterization (q_c) and for discrimination (q_d) using an adapted *F-Measure*, e.g., [12], resulting in the harmonic mean between q_c and q_d applying normalized quality functions with a value range in $[0; 1]$

$$F(q_c, q_d) = \frac{(1 + \beta^2) \cdot q_c \cdot q_d}{\beta^2 \cdot q_c + q_d}$$

Equivalently to the F-Measure used in information retrieval, $F(q_c, q_d)$ now measures the effectiveness of the concept description with respect to a user who attaches β times as much importance to q_c (characterization) as q_d (discrimination); for the

F-Measure $\beta = 1$, for equally weighting characterization and discrimination. The β parameter provides for a convenient option for adaptations, and for shifting the focus between characteristic and discriminative patterns.

In summary, we can view the concept description task using local patterns from different perspectives: From the *characterization* view, from the *discrimination view* and from a combined view using the F-Measure. We will discuss several practical options concerning concept characterization and discrimination in the next section when introducing the setup and results of the case study.

4 Case Study – Characterizing (Non-)Spammers

In this section, we discuss the results of the presented approach using a case study in the social bookmarking domain. We first introduce the context of the social bookmarking system BibSonomy. Next, we describe the applied data set and discuss the results of the application of the presented approach.

4.1 Bibsonomy

Resource sharing systems like BibSonomy provide an easy way to organize and manage different kinds of resources. What is considered as a resource depends on the type of system. For instance, in del.icio.us, the resources are URLs, in flickr, the resources are pictures, and in BibSonomy they are either URLs or publication entries. At their core, these systems are all very similar. Once a user is logged in, he can add a resource to the system, and assign arbitrary tags to it. The collection of all assignments of a user form the *personomy*, the collection of all personomies constitutes the *folksonomy*. As in other systems, the user can explore personomies of arbitrary users in all dimensions: For a given user one can see all resources that have been uploaded, together with the tags that were assigned to them; when clicking on a resource one can see which other users have uploaded this specific resource and how they tagged it; and when clicking on a tag one can see who assigned it to which resources [7]. Overall, these systems provide a very intuitive navigation through the data.

4.2 Features

In [15] we proposed 25 features in four categories to describe the users and their behaviour in BibSonomy. For our experiments we focused on the attributes and the categories that are not derived using information about spammers and non-spammers and can therefore be regarded as 'non-semantic' socio-demographic features. Therefore, we selected the 15 most interesting attributes (features) from three categories: Profile features, location-based features, and activity-based features. We shortly repeat these in this section for the convenience of the reader.

Profile features Table 1 shows profile features that are extracted from the profile of a user which he or she reveals when requesting an account in BibSonomy. Most of the

Feature name	Description
namedigit	name contains digits
namelen	length of name
maildigit	email address contains digits
maillen	length of mail address
realnamelen	length of realname
realnamedigit	realname contains digits
realname2	two realnames
realname3	three realnames

Table 1. Profile features

fields to fill in at registration are not obligatory, however, users need to indicate at least a name and a valid e-mail-address. In contrast to normal users, spammers often use names or e-mail addresses with many numbers. For instance, typical names of spammers are "styris888" and "painrelief2". The spam/non-spam distribution of the number of digits in the username, real name and the email address (*namedigit*, *realnamedigit*, *maildigit*) is very different. The *namelen*, *maillen* and *realnamelen* features refer to the length of the usernames and realnames in terms of characters. The *realname2* and *realname3* features are binary and set to one, if the user has indicated two or three names (legitimate users often register with their full names).

Location based features Location based features refer to describing the user's location and domain. Table 2 summarizes the location based features.

Often, the same spammer uses several accounts to publish the same content. These accounts show the same IP address when they are registered. Thus, if one user with a specific IP or uses a specific domain is already marked as a spammer, the probability that other users with the same IP or domain are also spammers is higher. When considering the users in the training dataset, we observed this for users of specific domains and created therefor the features (*domaincount, tldcount*). The probability that a user who is from a rare domain which hosts many spammers is also a spammer is higher than average (and vice versa). For instance, 16 users have registered with the domain "spambob.com" and 137 with the domain "rhinowebmail", all of which were classified as spammers.

Feature name	Description
tld	Top Level Domain
domaincount	number of users in the same domain
tldcount	number of users in the same top level domain

Table 2. Location based features

Activity based features Activity properties (Table 3) consider different kinds of user interaction with the social bookmarking system. While normal users tend to interact with the system instantly after their registration (e. g., by posting a bookmark), spam users often wait a certain time after they submit their first post. This timelag can be considered when characterizing spam (*datediff*).

There are other 'simple' properties which we found when manually cleaning the system from spam. For instance, '$group=public' is added by many spammers, since this specific tag is used by a software to generate spam in social bookmarking systems (*grouppub*). Furthermore, the number of tags per post often varies (*tasperpost*). Spammers usually add many different tags to a resource, either to show up more often when searching for many different tags, or to include 'good' tags in order to confuse spam detection mechanisms. Considering the BibSonomy dataset, spammers add in average eight tags to a post, while non-spammers add three. The average number of TAS (see definition in [16] is about 450 for spammers and 250 for users (*tascount*).

Feature name	Description
datediff	difference between registration and first post
grouppub	number of times '$group=public' was used
tasperpost	number of tags per post
tascount	number of total tags added to all posts of this account

Table 3. Activity based features

4.3 Results and Discussion

For the case study, we used data from the BibSonomy system, that is, the dataset provided by last years ECML PKDD discovery challenge.[1] The original data set contains 31715 cases (instances) in total. After removing instances with missing values, the applied data set contained 31034 instances. As discussed above, we applied the described 15 attributes for concept description. The distribution of the classes in the applied dataset is highly unbalanced, with 1812 non-spammers and 29222 spammers for default target shares of 5.8% non-spammers, and 94.2% spammers. In the following, we will discuss both classes, i.e., *spammers* and *non-spammers* as our target concepts, using both characteristic and discriminative features/subgroups.

For the spammer/non-spammer case study we applied the q_{TPR} quality function measuring the true positive rate, or the recall of the patterns. For the discriminative setting we applied the relative gain quality function q_{RG} which is order equivalent to precision. Finally, for assessing the F-Measure, we utilized the classical recall and precision measures. As outlined above, we used adapted measures including the *simplicity* of the patterns by favoring patterns with shorter descriptions.

[1] http://www.kde.cs.uni-kassel.de/ws/rsdc08/dataset.html

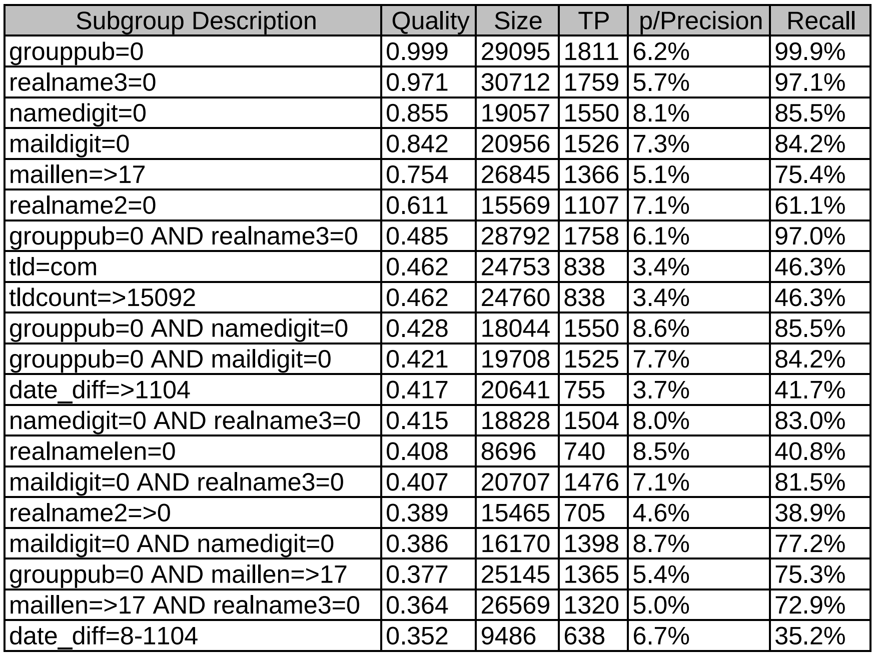
Since there exists a variety of other quality functions, especially for the disciminative setting, i.e., for class discrimination, we also experimented with other quality functions, e.g., the weighted relative accuracy (WRACC) funtion [17]: The relative gain measure obtains very specific (smaller) subgroups compared, e.g., to WRACC, which are very discriminative and focus on specific local points of the target space. Therefore, the selection and application of the quality functions – especially those used for discrimination – is significantly dependent on the user requirements and the goal parameters of interest that are to be included into the quality function: Since the criteria for the evaluation and classification of spammers of the case study already considered measures from information retrieval (precision and recall) both turned out to be the perfect candidates and could be directly included in the subgroup discovery techniques. In addition, the simplicity of the patterns was also perceived as a very important parameter for concept description.

In the following sections we present the results of applying the approach for describing spammers and non-spammers for concept characterization and discrimination.

Describing Non-Spammers: When comparing the attributes that are used for concept characterization and discrimination for non-spammers, we see that mainly date_diff, grouppub, maildigit, maillen, namedigit, realname2, realname3, realnamelen, tld, and tldcount are used for characterization, while date_diff, domaincount, maillen, namelen, realnamelen, tascount, tasperpost, tasperpost, tld, and tldcount are more discriminative. The used value ranges for the features are not always exclusive, which is explained by the general observation that characterizing features are also often observed for another class, while this is not true for the discriminative features. In general, profile information seems more important for characterization, while activity-based features seem more important for discrimination.

Figure 1 shows the results of applying the concept characterization task: It is easy to see, that the discovered subgroups are relatively large, and (by construction) large areas of the target space are covered by the individual patterns. The most important attributes are *grouppub*, *realname3*, and *namedigit* which characterize the non-spammer class with a value of zero very well. The latter observation can be confirmed by the fact that most spammers have numbers in their usernames while non-spammers focus on short usernames without numbers.

Figure 2 shows the results of applying the discrimination task: As expected, the discriminative setting focuses on relatively small sections of the target space with high precision (target share), in contrast to the concept characterization results. The resulting patterns are significantly discriminative for the target class (non-spammer) and are readily available, e.g., for classification or explanation. From the application point of view one may observe that the number of tags per post is a very important attribute. The most discriminative values for this attribute are 2 and 3 which better fits the intuition that the non-spammer adds a smaller number of tags to a resource than a spammer. However, 4 and 5 is still a discriminative number and appears again in a few patterns. Another very important attribute is *data_diff* with a value smaller than 7. Typically, non-spammers seem to register and submit their first post thereafter, while spammers tend to register in BibSonomy and wait until they start to use the service they 'recently' discovered on the



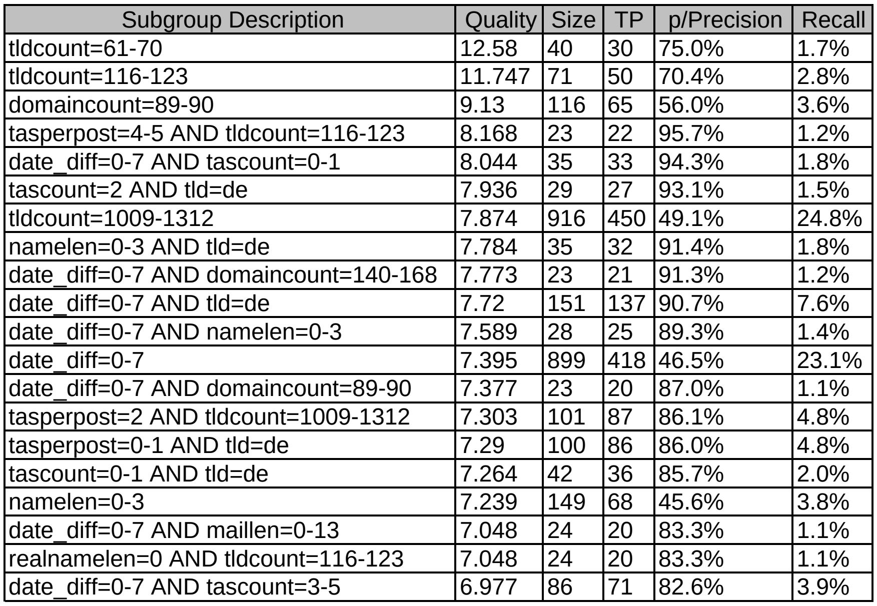
Subgroup Description	Quality	Size	TP	p/Precision	Recall
grouppub=0	0.999	29095	1811	6.2%	99.9%
realname3=0	0.971	30712	1759	5.7%	97.1%
namedigit=0	0.855	19057	1550	8.1%	85.5%
maildigit=0	0.842	20956	1526	7.3%	84.2%
maillen=>17	0.754	26845	1366	5.1%	75.4%
realname2=0	0.611	15569	1107	7.1%	61.1%
grouppub=0 AND realname3=0	0.485	28792	1758	6.1%	97.0%
tld=com	0.462	24753	838	3.4%	46.3%
tldcount=>15092	0.462	24760	838	3.4%	46.3%
grouppub=0 AND namedigit=0	0.428	18044	1550	8.6%	85.5%
grouppub=0 AND maildigit=0	0.421	19708	1525	7.7%	84.2%
date_diff=>1104	0.417	20641	755	3.7%	41.7%
namedigit=0 AND realname3=0	0.415	18828	1504	8.0%	83.0%
realnamelen=0	0.408	8696	740	8.5%	40.8%
maildigit=0 AND realname3=0	0.407	20707	1476	7.1%	81.5%
realname2=>0	0.389	15465	705	4.6%	38.9%
maildigit=0 AND namedigit=0	0.386	16170	1398	8.7%	77.2%
grouppub=0 AND maillen=>17	0.377	25145	1365	5.4%	75.3%
maillen=>17 AND realname3=0	0.364	26569	1320	5.0%	72.9%
date_diff=8-1104	0.352	9486	638	6.7%	35.2%

Fig. 1. Concept Characterization of *non-spammers*. The table shows top 20 subgroup descriptions for the target concept *class = non-spammer*; *Size* denotes the subgroup size, *Quality* is measured by the characteristic relative gain quality function, $p/Precision$ denotes the target share of the subgroup (precision), TP the number of *true positives* in the subgroup, and *Recall* the recall value of the subgroup pattern.

internet. Furthermore, the 'tld=de' features seems to be very important, which can be explained by the fact that the system is very popular (for legitimate) users in Germany.

In general, while the concept characterization tasks produces descriptions which focus on demographic features such as the selection of the username, a distinction between different groups of non-spammers can be made with a combination of demographic and activity features. We can learn from this, that a good indicator for non-spammers is already given in the data they provide when registering; however, in order to reliable classify spammers we need to add information about their interaction with the system.

Finally, Figure 3 shows the results of applying the combined F-Measure capturing both concept characterization and discrimination. Since the F-Measure combines both characterization and discrimination, the results show a balance between the other result tables: The focus of the patterns shifts towards more 'precise' but also more typical features. Considering the selected attributes, *date_diff* and *tldcount* appear most frequently. Considering the values of the attribute *tasperpost*, the attribute is still important; however, it only comprises a smaller number of tas (≤ 2), while the different subgroups implied by the condition tas (> 2) seem to form a bad general description.



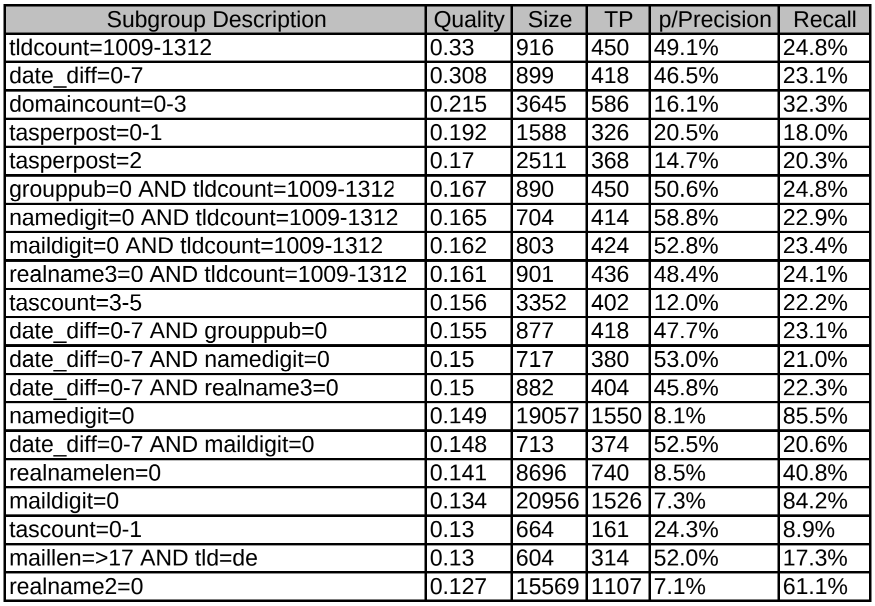
Subgroup Description	Quality	Size	TP	p/Precision	Recall
tldcount=61-70	12.58	40	30	75.0%	1.7%
tldcount=116-123	11.747	71	50	70.4%	2.8%
domaincount=89-90	9.13	116	65	56.0%	3.6%
tasperpost=4-5 AND tldcount=116-123	8.168	23	22	95.7%	1.2%
date_diff=0-7 AND tascount=0-1	8.044	35	33	94.3%	1.8%
tascount=2 AND tld=de	7.936	29	27	93.1%	1.5%
tldcount=1009-1312	7.874	916	450	49.1%	24.8%
namelen=0-3 AND tld=de	7.784	35	32	91.4%	1.8%
date_diff=0-7 AND domaincount=140-168	7.773	23	21	91.3%	1.2%
date_diff=0-7 AND tld=de	7.72	151	137	90.7%	7.6%
date_diff=0-7 AND namelen=0-3	7.589	28	25	89.3%	1.4%
date_diff=0-7	7.395	899	418	46.5%	23.1%
date_diff=0-7 AND domaincount=89-90	7.377	23	20	87.0%	1.1%
tasperpost=2 AND tldcount=1009-1312	7.303	101	87	86.1%	4.8%
tasperpost=0-1 AND tld=de	7.29	100	86	86.0%	4.8%
tascount=0-1 AND tld=de	7.264	42	36	85.7%	2.0%
namelen=0-3	7.239	149	68	45.6%	3.8%
date_diff=0-7 AND maillen=0-13	7.048	24	20	83.3%	1.1%
realnamelen=0 AND tldcount=116-123	7.048	24	20	83.3%	1.1%
date_diff=0-7 AND tascount=3-5	6.977	86	71	82.6%	3.9%

Fig. 2. Concept Discrimination of *non-spammers*. The table shows the 20 best subgroup descriptions for the target concept *class = non-spammer*, for the discrimination setting. We applied the quality function q_{RG}, i.e., the relative gain quality function; for a description of the remaining parameters see Figure 1.

Describing Spammers: Considering the attributes used for concept discrimination, we see that specific values of domaincount, grouppub, maildigit, namedigit, namelen, realname2, realnamelen, tasperpost, tldcount, and certain top-level domains (tld) are very good indicators for spammers. For characterization, attributes like date_diff, domaincount, grouppub, maildigit, maillen, namedigit, realnameX, tascount, tasperpost, tld and tldcount are important, similar to the discriminative setting, but as expected the value sets are often more general than the specific patterns used for discrimination.

Figure 4 shows the results of the characterization of spammers:[2] While *grouppub=>0* is a perfect feature for discrimination, *grouppub=0* is also a good feature for characterization, since there is also a large number of spammers with *grouppub=0*. As expected, spammers often do not enter very many names (*realname3=0*), but they tend to have long (*namelen=>9*) names, and a long email (*maillen=>17*). Additionally, spammers often use digits in their names and email (*namedigit, maildigit*) since it seems that they tend to number their created accounts at different sites. As a further characteristic, they often come from the *com* domain, and use main tas (*tascount=>33*) and tas per post as well (*tasperpost=5-11*).

[2] These results are also similar to the F-Measure results since spammers form the majority class and therefore the recall seems to dominate in this setting. Therefore we don't provide a detailed discussion of the F-Measure results.



Subgroup Description	Quality	Size	TP	p/Precision	Recall
tldcount=1009-1312	0.33	916	450	49.1%	24.8%
date_diff=0-7	0.308	899	418	46.5%	23.1%
domaincount=0-3	0.215	3645	586	16.1%	32.3%
tasperpost=0-1	0.192	1588	326	20.5%	18.0%
tasperpost=2	0.17	2511	368	14.7%	20.3%
grouppub=0 AND tldcount=1009-1312	0.167	890	450	50.6%	24.8%
namedigit=0 AND tldcount=1009-1312	0.165	704	414	58.8%	22.9%
maildigit=0 AND tldcount=1009-1312	0.162	803	424	52.8%	23.4%
realname3=0 AND tldcount=1009-1312	0.161	901	436	48.4%	24.1%
tascount=3-5	0.156	3352	402	12.0%	22.2%
date_diff=0-7 AND grouppub=0	0.155	877	418	47.7%	23.1%
date_diff=0-7 AND namedigit=0	0.15	717	380	53.0%	21.0%
date_diff=0-7 AND realname3=0	0.15	882	404	45.8%	22.3%
namedigit=0	0.149	19057	1550	8.1%	85.5%
date_diff=0-7 AND maildigit=0	0.148	713	374	52.5%	20.6%
realnamelen=0	0.141	8696	740	8.5%	40.8%
maildigit=0	0.134	20956	1526	7.3%	84.2%
tascount=0-1	0.13	664	161	24.3%	8.9%
maillen=>17 AND tld=de	0.13	604	314	52.0%	17.3%
realname2=0	0.127	15569	1107	7.1%	61.1%

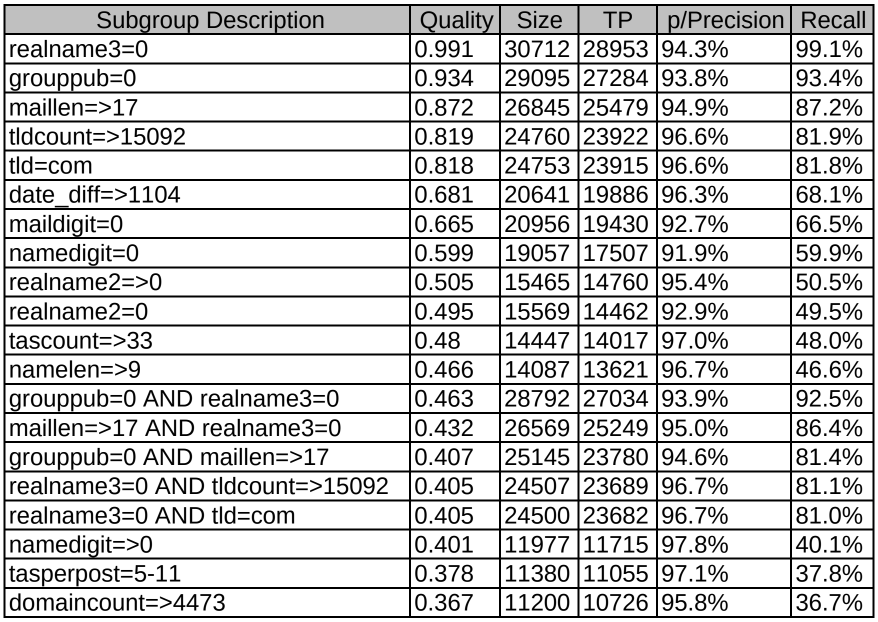
Fig. 3. Concept Description using the F-Measure. The table shows the top 20 subgroup descriptions for the target concept *class* = *non-spammer* (combined concept description setting).

Figure 5 shows the results of the discriminative description of spammers: While *date_diff* is not so important for discriminating spammers than discriminating non-spammers, the tasperpost attribute is also very important. As expected, and as also shown by the characterization findings, realnamelen, maildigit, namedigit provide typical features for spammers – usually having digits in their names, and using longer names in general. Another very discriminative feature is *tld*. Spammer seems to heavily rely on domains like: *th*, *us*, *info*,and *biz* in addition to the already mentioned *com* domain. This complements the patterns observed for the non-spammers.

5 Related Work

This paper is especially interesting with respect to two research areas: From a theoretical point of view, we propose a novel approach for describing concepts based on subgroup discovery methods. Practically, we apply our method to the application of spam detection in social bookmarking systems. In this section we will discuss related work for both parts.

Several methods for concept description have been investigated in the past: For example, attribute-oriented induction techniques, e.g., [1] generate a set of generalized relations for obtaining a summary of the task-relevant data. Attribute-oriented induction focuses on a specific concept of interest, relies on the specification of the relevant set of attributes and applies concept hierarchies for concept generalization.



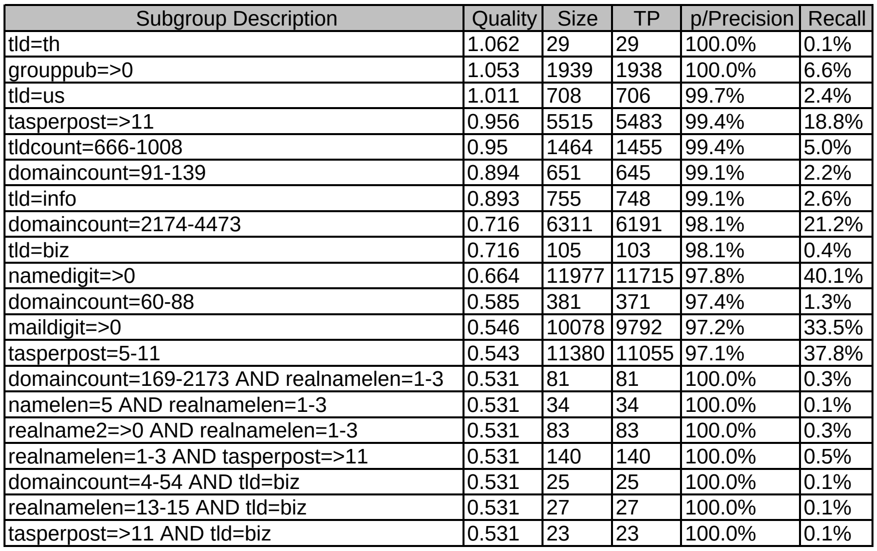
Subgroup Description	Quality	Size	TP	p/Precision	Recall
realname3=0	0.991	30712	28953	94.3%	99.1%
grouppub=0	0.934	29095	27284	93.8%	93.4%
maillen=>17	0.872	26845	25479	94.9%	87.2%
tldcount=>15092	0.819	24760	23922	96.6%	81.9%
tld=com	0.818	24753	23915	96.6%	81.8%
date_diff=>1104	0.681	20641	19886	96.3%	68.1%
maildigit=0	0.665	20956	19430	92.7%	66.5%
namedigit=0	0.599	19057	17507	91.9%	59.9%
realname2=>0	0.505	15465	14760	95.4%	50.5%
realname2=0	0.495	15569	14462	92.9%	49.5%
tascount=>33	0.48	14447	14017	97.0%	48.0%
namelen=>9	0.466	14087	13621	96.7%	46.6%
grouppub=0 AND realname3=0	0.463	28792	27034	93.9%	92.5%
maillen=>17 AND realname3=0	0.432	26569	25249	95.0%	86.4%
grouppub=0 AND maillen=>17	0.407	25145	23780	94.6%	81.4%
realname3=0 AND tldcount=>15092	0.405	24507	23689	96.7%	81.1%
realname3=0 AND tld=com	0.405	24500	23682	96.7%	81.0%
namedigit=>0	0.401	11977	11715	97.8%	40.1%
tasperpost=5-11	0.378	11380	11055	97.1%	37.8%
domaincount=>4473	0.367	11200	10726	95.8%	36.7%

Fig. 4. Concept Characterization of *spammers*. The table shows 20 best subgroup descriptions for the target concept *class* = *spammer*; as for the non-spammer characterization, we applied the true positive rate q_{TPR} quality function; for a description of the parameters see Figure 1.

Furthermore, association rule mining [9] can also be regarded as a general concept description task. However, in contrast to subgroup discovery and attribute-oriented induction, association rule algorithms do not focus on a specific target concept. In relation to to mining association rules, the subgroup discovery task is focused on a specific concept of interest. Therefore, less (non-interesting) results are generated, and efficient algorithms can be applied that utilize the concept of interest for pruning the search space [8]. Furthermore, the proposed approach condenses the discovered set of subgroups into a set of relevant subgroups that still convey the same information as the original set.

Compared to attribute-oriented induction subgroup discovery can especially be applied for a large number of independent variables/attributes of interest; it can integrate background knowledge [18] for decreasing the search space and to focus the search process. Thus, it can support much larger search spaces. The discovery process is goal-directed and can be flexibly configured using an arbitrary quality function: Then, during the description task, the quality function is directly applied for guiding the search (generalization) process, in contrast to attribute-oriented techniques.

Mining contrast sets, e.g., [19], focuses on discovering conjunctions of attribute-value pairs that differ meaningfully between groups (concepts) concerning their distributions. In that sense, subgroup discovery for binary target concepts can be regarded as a special case of constrast set mining, however, Kralj et al. have inductively shown that subgroup discovery and contrast set mining are actually compatible approaches. There-



Subgroup Description	Quality	Size	TP	p/Precision	Recall
tld=th	1.062	29	29	100.0%	0.1%
grouppub=>0	1.053	1939	1938	100.0%	6.6%
tld=us	1.011	708	706	99.7%	2.4%
tasperpost=>11	0.956	5515	5483	99.4%	18.8%
tldcount=666-1008	0.95	1464	1455	99.4%	5.0%
domaincount=91-139	0.894	651	645	99.1%	2.2%
tld=info	0.893	755	748	99.1%	2.6%
domaincount=2174-4473	0.716	6311	6191	98.1%	21.2%
tld=biz	0.716	105	103	98.1%	0.4%
namedigit=>0	0.664	11977	11715	97.8%	40.1%
domaincount=60-88	0.585	381	371	97.4%	1.3%
maildigit=>0	0.546	10078	9792	97.2%	33.5%
tasperpost=5-11	0.543	11380	11055	97.1%	37.8%
domaincount=169-2173 AND realnamelen=1-3	0.531	81	81	100.0%	0.3%
namelen=5 AND realnamelen=1-3	0.531	34	34	100.0%	0.1%
realname2=>0 AND realnamelen=1-3	0.531	83	83	100.0%	0.3%
realnamelen=1-3 AND tasperpost=>11	0.531	140	140	100.0%	0.5%
domaincount=4-54 AND tld=biz	0.531	25	25	100.0%	0.1%
realnamelen=13-15 AND tld=biz	0.531	27	27	100.0%	0.1%
tasperpost=>11 AND tld=biz	0.531	23	23	100.0%	0.1%

Fig. 5. Concept Discrimination of *spammers*. The table shows the top 20 subgroup descriptions for the target concept *class = spammer*, for the discrimination setting. We applied the quality function q_{RG}, i.e., the relative gain quality function; for a description of the remaining parameters see Figure 1.

fore, the presented pattern discovery task provides for more options: It is more flexible concerning the application requirements, weighting recall and precision with respect to the characterization and discrimination tasks.

Research on spam detection in social media has been conducted by the blog and wikipedia community. Methods to detect comment spam and spam blogs have been proposed by [20–22]. [23, 24] are the first to deal with spam in tagging systems explicitly. The authors identify anti-spam strategies for tagging systems and construct and evaluate models for different tagging behaviour. In contrast to [23, 24] we present a concrete study using data mining techniques to get more insights on a real-world dataset. [15] focuses on the extraction of features suitable to predict spam behavior in social bookmarking systems by utilizing machine learning approaches.

6 Conclusions

In this paper, we have presented an approach for concept characterization and discrimination using local patterns that are discovered using subgroup discovery techniques. We have shown, how to apply suitable quality functions for the characterization and discrimination tasks. Additionally, we presented an option for weighting these sub-goals in order to provide a combined option for a comprehensive view on the the characterization and discrimination aspects. These can then be analyzed according to the specific requirements of the application.

We described a case study using real-world data from the BibSonomy system that provided for a very versatile testbed. The presented approaches reveal interesting insights about the specific characteristics of the BibSonomy (non)spammers. For example, the number of tags per post or the time lag between registration and the first post help to distinguish spammers from (non)spammers.

For future work, we plan to extend the case study by comparing different profiles that were involved in tagging spammers. Then, we can easily perform an assessment of the inter-annotator agreement of these users. A next step is the extension of the discovery technique for community mining in social networks by considering the special (triadic) structure of the user-tag-resource space. Additionally, we plan to integrate approaches for optimizing the set of patterns, e.g., similar to pattern teams [25]. Another interesting option for future work is given by considering extended measures for objective and subjective interestingness criteria besides a potentially more refined simplicity/complexity measure.

Acknowledgements

This work has been partially supported by the German Research Council (DFG) under grant Pu 129/8-2 and grant GZ STU 252/5-1.

References

1. Han, J., Fu, Y.: Exploration of the Power of Attribute-Oriented Induction in Data Mining. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., eds.: Advances in Knowledge Discovery and Data Mining. AIII Press/MIT Press (1996)
2. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., eds.: Advances in Knowledge Discovery and Data Mining. AAAI Press (1996) 249–271
3. Wrobel, S.: An Algorithm for Multi-Relational Discovery of Subgroups. In: Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97), Berlin, Springer Verlag (1997) 78–87
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd Edition. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2006)
5. Lemmerich, F., Atzmueller, M.: Incorporating Exceptions: Efficient Mining of ϵ-Relevant Subgroup Patterns. Proc. LeGo-09 Workshop at ECML/PKDD 2009 (2009)
6. Atzmueller, M., Puppe, F.: Semi-automatic refinement and assessment of subgroup patterns. In: Proc. 21th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2008). (2008) 323–328
7. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: BibSonomy: A social bookmark and publication sharing system. In: CS-TIW '06, Aalborg, Denmark, Aalborg University Press (July 2006)
8. Atzmueller, M., Puppe, F.: SD-Map – A Fast Algorithm for Exhaustive Subgroup Discovery. In: Proc. 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006). Number 4213 in LNAI, Berlin, Springer Verlag (2006) 6–17
9. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In Bocca, J.B., Jarke, M., Zaniolo, C., eds.: Proc. 20th Int. Conf. Very Large Data Bases, (VLDB), Morgan Kaufmann (1994) 487–499

10. Garriga, G.C., Kralj, P., Lavrač, N.: Closed Sets for Labeled Data. Journal of Machine Learning Research **9** (2008) 559–580
11. Lavrac, N., Gamberger, D.: Relevancy in Constraint-based Subgroup Discovery. In Jean-Francois Boulicaut, Luc de Raedt, H.M., ed.: Constraint-based Mining and Inductive Databases. Volume 3848 of LNCS. Springer Verlag, Berlin (To appear, 2006)
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press (2008)
13. Atzmueller, M., Baumeister, J., Puppe, F.: Quality Measures and Semi-Automatic Mining of Diagnostic Rule Bases. In: Proc. 15th Intl. Conference on Applications of Declarative Programming and Knowledge Management (INAP). Volume 3392 of LNAI., Berlin, Springer Verlag (2005) 65–78
14. Koenig, R., Johansson, U., Niklasson, L.: Using Genetic Programming to Increase Rule Quality. In: Proc. 21st Intl. Florida Artificial Intelligence Research Society Conference (FLAIRS), AAAI Press (2008) 288–293
15. Krause, B., Schmitz, C., Hotho, A., Stumme, G.: The anti-social tagger - detecting spam in social bookmarking systems. In: Proc. AIRWeb '08. (2008)
16. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: Proc. ESWC '06, Budva, Montenegro, Springer (June 2006) 411–426
17. Lavrac, N., Kavsek, B., Flach, P., Todorovski, L.: Subgroup Discovery with CN2-SD. Journal of Machine Learning Research **5** (2004) 153–188
18. Atzmueller, M., Puppe, F., Buscher, H.P.: Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In: Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05), Edinburgh, Scotland (2005) 647–652
19. Kralj, P., Lavrac, N., Webb, G.I.: Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set and Emerging Pattern and Subgroup Mining. Journal of Machine Learning Research **10** (2009) 377–403
20. Kolari, P., Finin, T., Joshi, A.: SVMs for the Blogosphere: Blog Identification and Splog Detection. AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (2006)
21. Kolari, P., Java, A., Finin, T., Oates, T., Joshi, A.: Detecting Spam Blogs: A Machine Learning Approach. AAAI '06 (2006)
22. Mishne, G., Carmel, D., Lempel, R.: Blocking blog spam with language model disagreement. In: Proc. AIRWeb '05, New York, NY, USA, ACM (2005) 1–6
23. Heymann, P., Koutrika, G., Garcia-Molina, H.: Fighting spam on social web sites: A survey of approaches and future challenges. IEEE Internet Computing **11**(6) (2007) 36–45
24. Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems. In: Proc. AIRWeb '07, New York, NY, USA, ACM (2007) 57–64
25. Knobbe, A.J., Ho, E.K.Y.: Pattern Teams. In: Knowledge Discovery in Databases: Proc. PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Volume 4213 of LNCS., Springer (2006) 577–584