Tags vs Shelves:
From Social Tagging to Social Classification

Arkaitz Zubiaga*
NLP & IR Group @ UNED
Juan del Rosal, 16
28040 Madrid, Spain
azubiaga@lsi.uned.es

Christian Körner*
Knowledge Management Institute
Graz University of Technology
christian.koerner@tugraz.at

Markus Strohmaier
Knowledge Management Institute
Graz University of Technology
and Know-Center
markus.strohmaier@tugraz.at

ABSTRACT

Recent research has shown that different tagging motivation and user behavior can effect the overall usefulness of social tagging systems for certain tasks. In this paper, we provide further evidence for this observation by demonstrating that tagging data obtained from certain types of users - so-called Categorizers - outperforms data from other users on a social classification task. We show that segmenting users based on their tagging behavior has significant impact on the performance of automated classification of tagged data by using (i) tagging data from two different social tagging systems, (ii) a Support Vector Machine as a classification mechanism and (iii) existing classification systems such as the Library of Congress Classification System as ground truth. Our results are relevant for scientists studying pragmatics and semantics of social tagging systems as well as for engineers interested in influencing emerging properties of deployed social tagging systems.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms

Algorithms, Human Factors, Measurement

Keywords

Tagging, Folksonomies, Classification, Libraries

1. INTRODUCTION

Recent research on social tagging systems has in part been motivated by a vision that the data produced by users (so-called taggers) can be used for social classification, i.e., the collective classification of resources into a commonly agreed structure. While libraries and librarians have performed the task of classification for centuries, the process of manually categorizing resources is expensive. The Library of Congress in the United States for example reported that the average cost of cataloging a bibliographic record by professionals was $94.58 in 2002[1]. Given these costs, social classification systems and algorithms represent an interesting alternative. Social tagging systems like Delicious[2], LibraryThing[3] or GoodReads[4] have demonstrated their ability to quickly generate large amounts of metadata in the form of tags. These tags have been shown to be useful for, for example, information access and organization. It has also been shown that social tags outperform traditional content-based approaches in many cases for tasks like information retrieval [8] and automated classification [28]. Yet, little is known about the usefulness of social tagging data for classifying resources, or about the type of tagging behavior that yields the best classification results.

The effectiveness of tagging data has been found to differ among different user populations and tasks [13]. In this work, we build on two existing distinctions of tagging motivation - *Describers* and *Categorizers* - introduced in our previous work [23] and further elaborated in [13]. According to this distinction, some users use tags to describe resources (Describers), while others use tags to categorize them (Categorizers). In past research, it has been shown that tags produced by Describers are *superior* for certain tasks such as information retrieval [8] or knowledge acquisition [12]. In this paper, we report on a task where descriptive tags seem *inferior*. To the best of our knowledge, this paper is the first to (i) identify social classification as a task where descriptive tags seem inferior and (ii) confirm that Categorizers outperform Describers on this task. Our results further advance previous research suggesting that user behavior (the pragmatics of tagging) influences the effectiveness of tagging data for different tasks.

To this end, we perform a set of descriptiveness and classification experiments with both Describers and Categorizers on two social tagging systems for books: LibraryThing and GoodReads. We analyze how tags by each kind of users

*Both authors contributed equally to this work

HT'11, June 6–9, 2011, Eindhoven, The Netherlands.

[1] http://www.loc.gov/loc/lcib/0302/collections.html
[2] http://delicious.com
[3] http://www.librarything.com
[4] http://www.goodreads.com

resemble to (1) descriptions of books, and (2) expert-driven categorization of books. Our results confirm that differences in tagging behavior exist, and that users who provide fewer descriptive tags (i.e., Categorizers) perform better for the classification task.

The paper is structured as follows: In Section 2, we introduce the characteristics of social tagging systems and the related terminology. Section 3 reviews and presents related work. In Section 4, we introduce selected aspects of user motivation in social tagging systems, and we detail some measures that can be used to identify them. Then, in Section 5, we describe the settings of our experiments, analyzing their results in Section 6. Finally, we conclude the paper in Section 7, and outlook on future work in Section 8.

2. TERMINOLOGY

Social tagging systems allow users to save and annotate resources (e.g., web pages, movies or books) with freely chosen, optional words - so called *tags* - and share them with the community. Saving and annotating such resources helps users maintaining a collection of their resources of interest, in such a way that enables searching and accessing them by taking advantage of annotated tags. All these annotations are said to be social when they are shared with the community. The tag structure resulting from community's annotations makes possible to apply algorithms in order to create a so-called folksonomy. Folksonomy is a neologism, a portmanteau of *folk* (people), *taxis* (classification) and *nomos* (management), in other words a classification managed by people. Usually folksonomies are represented by tripartite graphs with hyper edges. These structures contain three finite, disjoint sets which are 1) a set of users $u \in U$, 2) a set of resources $r \in R$ and 3) a set of tags $t \in T$ annotating resources R. A folksonomy as a whole is defined as the annotations $F \subseteq U \times T \times R$ (cf. [17]). Subsequently a personomy of a user $u \in U$ is the reduction of a folksonomy F to the user u [9]. In the following a *tag assignment* (tas = (u,t,r); $tas \in TAS$) is a specific triple of one user $u \in U$, one tag $t \in T$ and one resource $r \in R$. A *bookmark* or *post* refers to a single resource r and all corresponding tags t of a user u. See Figure 1 for an example of a folksonomy of a social tagging system.

Not all tagging systems operate in the exact same way though. Certain social tagging systems impose certain constraints, e.g., by setting who is able to annotate which resource in what way. In this sense, two kinds of tagging systems can be distinguished [22]:

- **Simple tagging:** users describe their own resources, such as photos on Flickr.com, news on Digg.com or videos on Youtube.com, but nobody else annotates others' resources. Usually, the author of the resource is who annotates it. This means no more than one user tags a resource. The purpose of tags of these systems is primarily the improving of search and retrieval for others.

- **Collaborative tagging:** many users annotate the same resource, and all of them can tag it with tags in their own vocabulary. The collection of tags assigned by a single user creates a smaller folksonomy, also known as personomy. As a result, several users tend to post the same resource. For instance, CiteULike.org, LibraryThing.com and Delicious are based on collaborative annotations, where each resource (papers, books and URLs, respectively) can be annotated and tagged by all the users who consider it interesting.

This work focuses on social tagging systems with a collaborative perspective. Unlike simple tagging systems, they give the opportunity to further explore the aggregated annotations on each resource, and to analyze whether some of those annotations are more useful when it comes to classifying resources.

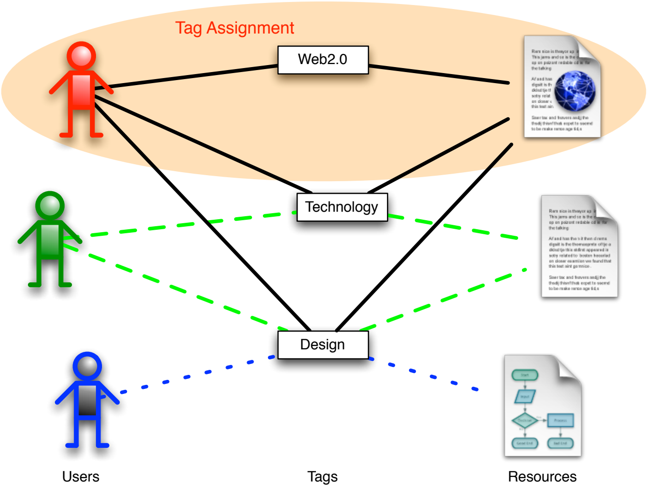

Figure 1: A folksonomy comprising users, tags and resources. A bookmark refers to all tags a user applies to a given resource.

3. RELATED WORK

Two topics are relevant in the context of this work: analysis of user behavior in social tagging systems, as well as the exploitation of annotations from these sites for the sake of automated classification tasks.

3.1 User Behavior

The first influential study on the topic of user behavior in social tagging systems is by Golder et al. [4]. This work analyzes the structure of such systems and the activity of the users, and presents a dynamic model of social tagging. Heckner et al. [7] examine the usage of tags in four different social tagging systems and explore how the resource type influences the tag choice and their usage within these systems (e.g. videos and photos are tagged more extensively than research articles). In another work, Chi et al. [1] study the efficiency of tags in tagging systems with the help of information theory. Their results show that the effectiveness of tags to refer to individual objects is waning. Wash et al. [24] interview users of Delicious in order to gain information about the incentives of the users in tagging systems. The main reasons they find are later retrieval, sharing and social recognition, among others. In another work by Rader et al. [20], the authors analyze the influences on tag choices in the popular social tagging system Delicious. One of the results of this work is that users' tag choice is driven by personal management in contrast to contributing to a shared vocabulary. Lipczak et al. [14] analyze the role of a resource's title for the selection of the resource's tags. In this work the authors show that, given two words with the same meaning, users tend to choose the tag which is also found in the title. However, an interesting finding is that, despite the tendency towards the title, users focus on maintaining consistency within their own profile.

With regard to our previous work, we have studied two different types of tagging motivation - Categorizers and Describers. In [23], we introduce these types of user motivation and give an overview on how tagging motivation varies across and within folksonomies. Furthermore, an outlook is given on how the variety of motivation in such systems can affect resulting folksonomies. In [13], we evaluate different measures to separate these types of users in both qualitative and quantitative ways, and show that tagging motivation can be approximated with simple statistical measures. In a subsequent paper [12], we study the influence of behavior and motivation on the semantic structure resulting from a folksonomy by showing that more verbose taggers are better for the identification of synonyms.

Building on this line of work, the presented paper focuses on studying the influence of user behavior and motivation on a different task: social classification.

3.2 Classification

There is little work dealing with the analysis of the usefulness of social tags for classification tasks. An early work by Noll and Meinel [18] presents a study of the characteristics of social annotations provided by end users, in order to determine their usefulness for web page classification. The authors matched user-supplied tags of a page against its categorization by the expert editors of the Open Directory Project (ODP). They evaluated at which hierarchy depth matches occurred, concluding that tags may perform better for broad categorization of documents rather than for more specific categorization. Also, Noll and Meinel [19] studied three types of metadata about web documents: social annotations (tags), anchor texts of incoming hyperlinks, and search queries to access them. They conclude that tags are better suited for classification purposes than anchor texts or search keywords.

In our previous work, we presented a study on the use of social annotations for web page classification, applied to the ODP categorization scheme [28]. We studied several approaches for representing tags in a vector space model in search of the optimal SVM classification accuracy results. On the one hand, we analyzed if considering the full tag set of each resource is helpful, or a subset of top tags should rather be considered. On the other hand, we also analyzed whether or not the number of users who assign each tag should be used as a weight representing the term frequency. Our study suggests considering all the tags and keeping their weights according to the number of users.

In another work where social tags were exploited for the benefit of web page classification, Godoy and Amandi [3] also showed the usefulness of social tags for web page classification, which outperformed classifiers based on full-text of documents. Going further, they concluded that stemming the tags reduces the performance of such classification, even thought some operations such as removal of symbols, compound words and reduction of morphological variants have a discrete positive impact on the task.

With regard to the classification of resources other than web pages, Lu et al. [15] present a comparison of tags annotated on books and their Library of Congress subject headings. Actually, no classification experiments are performed, but a statistical analysis of the tagging data shows encouraging results. By means of a shallow analysis of the distribution of tags across the subject headings, they conclude that user-generated tags seem to provide an opportunity for libraries to enhance the access to their resources. In addition, social tags have been used for clustering. In Ramage et al. [21], the inclusion of tagging data improved the performance of two clustering algorithms when compared to content-based clustering. This paper found that tagging data is more effective for specific collections than for a collection of general documents.

The unique idea of this paper is to bring together research on tagging behavior with research on classification algorithms in order explore (i) to what extent tagging data can be used for social classification tasks and (ii) whether certain user behavior yields better performance on this task.

	Categorizer	*Describer*
Goal of Tagging	later browsing	later retrieval
Change of Tag Vocabulary	costly	cheap
Size of Tag Vocabulary	limited	open
Tags	subjective	objective

Table 1: Two Types of Taggers

4. IDENTIFYING USER BEHAVIOR

As an approach to discriminate users by their behavior, we rely on a differentiation we established in previous works such as [12, 13, 11] - the notion of *Categorizers* and *Describers*.

Early works such as [16, 5] and [6] suggest that a distinction between at least two types of user motivations for tagging is interesting: on one hand, users can be motivated by categorization (in the following called *Categorizers*). These users view tagging as a means to categorize resources according to some (shared or personal) high-level conceptualizations. They typically use a rather elaborated tag set to construct and maintain a navigational aid to the resources for later browsing. In the context of libraries, one could think of Categorizers as those user who rely on a shelf-driven perspective in their annotations, as librarians would do when cataloging books. On the other hand, users who are motivated by description (so called *Describers*) view tagging as a means to accurately and precisely detail resources. These users tag because they want to produce annotations that are useful for later search and retrieval. The development of a personal, consistent ontology to navigate across their resources is not their intuition. Table 1 gives an overview of characteristics of the two different types of users, based on [11].

4.1 Measures

We use three different measures to differentiate users into Categorizers and Describers: Tags Per Post (TPP), Tag Resource Ratio (TRR), and Orphan Ratio (ORPHAN). In [13] additional measures are shown, however due to the high correlation between the measures in this paper and the measures presented additionally in [13] we limited our efforts to the ones detailed below. These measures rely on two features of user behavior: verbosity, which measures the number of tags a user tends to use when annotating, and diversity, which measures the extent to which users are using new tags that were not applied by themselves earlier. It is worthwhile noting that these measures provide one value for each user. The measure corresponding to each user is thus computed by considering the characteristics of her bookmarks and attached tag assignments. The resulting measures are then ranked in a list along with the rest of the users. This list makes possible inferring to what extent a user is rather a Categorizer or a Describer.

4.1.1 Tags per Post (TPP)

As a Describer would focus on describing her resources in a very detailed manner, the number of tags used to annotate each resource can be taken into account as an indicator to identify the motivation of the analyzed user. The *tags per post* measure (short TPP) captures this by dividing the number of all tag assignments of a user by the number of resources (see Equation 1). T_{ur} is the number of tags annotated by user u on resource r, and R_u is the number of resources of a user u. The more tags a user utilizes to annotate the resources the more likely she is a Describer and this would reflect in a higher TPP score.

$$TPP(u) = \frac{\sum^{r} |T_{ur}|}{|R_u|} \tag{1}$$

This measure relies on the verbosity of users, as it computes the average number of tags they assigned to bookmarks.

4.1.2 Orphan Ratio (ORPHAN)

Since Describers do not have a fixed vocabulary and freely choose tags to describe their resources in a detailed manner, they would not focus on reusing tags. This factor is analyzed in the *orphan ratio* (short $ORPHAN$). This measure relates the number of seldom used tags to the total number of tags. Equation 2 shows how seldom used tags are defined by the individual tagging style of a user. In this equation, t_{max} denotes the tag which was used the most by the user. Equation 3 shows the calculation of the final measure where T_u^o are seldom used tags and T_u are all tags of the given user. The more seldom used tags a user has the higher the orphan ratio is and the more she is a describer.

$$n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \tag{2}$$

$$ORPHAN(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t||R(t)| \leq n\} \tag{3}$$

By measuring whether users frequently use the same tags or rather rely on new ones, the ORPHAN ratio considers their diversity.

4.1.3 Tag Resource Ratio (TRR)

The *tag resource ratio* (short TRR) relates the number of tags of a user (i.e., the size of her vocabulary) to the total number of annotated resources (see Equation 4). A typical Categorizer would apply only a small number of tags to her resources and therefore score a low number on this measure.

$$TRR(u) = \frac{|T_u|}{|R_u|} \tag{4}$$

This measure relies on both verbosity, because users who use more tags in each bookmark would usually result in a higher TRR value, and diversity, as those who frequently use new tags will have a larger vocabulary. Nonetheless, the latter has a higher impact in this case, since the former could be altered by verbose users who tend to reuse tags.

5. EXPERIMENTS

This section presents the datasets used as well as the setting of our experiments.

5.1 Datasets

We use two social tagging sites of books for this work: *LibraryThing* and *GoodReads*. Both of them have a rather large community, and a large collection of annotated books. As of January 2011, LibraryThing has more than 1.2m users[5], whereas GoodReads has about 3.5 million users as of November 2010[6]. First, we queried the two sites for popular resources. We consider a resource to be popular if at least 100 users have annotated it as a bookmark, since it was shown that the tag set of a resource tends to converge when that many users contribute to it [4]. This way, we found an intersection of 65,929 popular books. Next, we looked for classification labels assigned by experts to this set of books. For this purpose we fetched their classification for both the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC) systems. The former is a classical taxonomy that is still widely used in libraries, whereas the latter is used by most research and academic libraries. We found that 27,299 books were categorized on DDC, and 24,861 books have an LCC category assigned to it. In total, there are 38,149 books with category data from either one or both category schemes. For the experiments, we rely on the first level of these classification schemes. At this level, DDC is made up by 10 categories, whereas LCC comprises 21 categories. For the latter, though, we reduce the number of categories to 20 - we merged E (*History of America*) and F (*History of the United States and British, Dutch, French, and Latin America*) categories into a single one, as it is not clear that they are disjoint categories.

Finally, we queried LibraryThing and GoodReads for gathering all the personomies (i.e., the whole collection of bookmarks and annotations of a given user) involved in the set of categorized books. Both sites present no restrictions on the bookmarks shown in personomies, so that they return all available public bookmarks for the queried users. At the time of fetching personomies, we got the full list of the bookmarks for each user. Each bookmark includes the user who saved it, an identifier of the annotated book, and a set of tags the user attached to it. In this process, we saved all the tags attached to each bookmark, except for GoodReads. In this case, a tag is automatically attached to each bookmark depending on the reading state of the book: *read*, *currently-reading* or *to-read*. We do not consider this to be part of the tagging process, but just an automated step, and we removed all their appearances in our dataset. Also, attaching tags to a bookmark is an optional step, so that depending on the social tagging site, a number of bookmarks may remain without tags. Table 2 presents the number of users, book-

[5] `http://www.librarything.com/users.php`

[6] `http://nospinpr.com/2010/11/22/goodreads-for-authors/`

marks and resources we gathered for each of the datasets, as well as the percent with attached annotations. In this work, as we rely on tagging data, we only consider annotated data, ruling out bookmarks without tags. Thus, from now on, all the results and statistics presented are based on annotated bookmarks.

LibraryThing			
	Annotated	Total	Percent
Users	153,606	400,336	38.37%
Bookmarks	22,343,427	44,612,784	50.08%
Resources	3,776,320	5,002,790	75.48%
Tags		2,140,734	-
GoodReads			
	Annotated	Total	Percent
Users	110,344	649,689	16.98%
Bookmarks	9,323,539	47,302,861	19.71%
Resources	1,101,067	1,890,443	58.24%
Tags		179,429	-

Table 2: Statistics on availability of tags in users, bookmarks, and resources for the three datasets.

Besides tagging data, we also gathered a set of descriptive data for each book from other sites. Since we do not have access to the books' content itself, we consider other sources for the descriptive data. These data include the following:

- Synopsis from Barnes & Noble[7]: a brief summary of the content of a book.
- User reviews from LibraryThing, GoodReads and Amazon[8]: comments provided by users on these sites for each book.
- Editorial reviews from Amazon: summaries written by experts.

Summarizing, our dataset comprises a set of books. Each record includes (i) a set of bookmarks, which have the form of a triple of user, book, and tags, (ii) synopses and reviews representing their description, and (iii) categorization data by experts.

5.2 Experimental Setup

The main objective of our work is to analyze how different sets of users are contributing to the classification or descriptiveness of the resources. According to the measures introduced above, we get ranked lists of users, where Categorizers rank high, and Describers rank low (this is arbitrary and could be inverted as well). With that, we select a set of users in the top as Categorizers, and another set in the tail as Describers. Both sets should have the same size in order to compare them. With these two sets, we perform classification and descriptiveness experiments to know how suitable they are for each of the tasks.

Figure 2 shows the distribution of the three measures we calculated for users on both social cataloging systems. The x axis represents quantiles of values, whereas y axis represents the number of users belonging to each quantile. The plots are quite similar in this case for both book datasets. TPP is the measure that requires more Categorizers, as compared to the number of Describers, to reach the same number of tag assignments in a given percent. This seems obvious, because TPP relies on verbosity. Next, ORPHAN also requires a larger number of Categorizers than Describers. To a lesser extent, though. And last, the opposite happens with TRR, since it requires larger number of Describers than Categorizers for the same percents. There is no reason that the last two measures have to yield on larger number of Categorizers, as they do not exclusively rely on verbosity, but mainly on diversity.

To choose the sets of users to perform the experiments with, we split the ranked lists by getting some of the top and bottom users. Choosing fixed percents of users would be unfair, though. Some users are likely to be more verbose, by definition, and they usually provide much more tag assignments than others. Thus, we split the users according to the percent of tag assignments they provide[9]. This enables a fairer split of the users, with the same amount of data, e.g., a 10% split ensures that both sets include 10% of all tag assignments, but the number of users differs among them. Figure 3 shows an example of how splitting by number of tag assignments can differ from splitting by number of users. With regard to the application of this splitting method in our datasets, using the three studied measures, Figure 4 gives a detailed overview of the results, showing percentages and the corresponding number of users in the subsets.

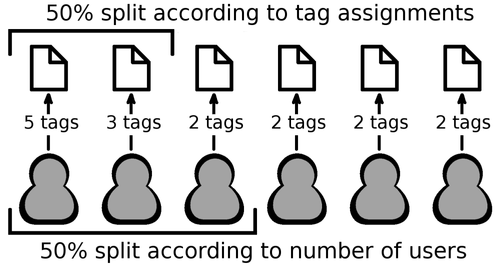
50% split according to tag assignments
5 tags
3 tags
2 tags
2 tags
2 tags
2 tags
50% split according to number of users

Figure 3: Example of a 50% segment by splitting based on tag assignments or number of users. Splitting by number of users is unfair, since it may yield bigger amounts of data.

5.2.1 Tag-based Classification

By tag-based classification, we consider the task of automatically assigning each book a category of the taxonomy by taking advantage of tagging data. This enables comparing tags provided by users to the categorization made by experts. Regarding the algorithm we use for the classification tasks, we rely on our previous work on the analysis of multi-class SVMs [27]. We analyzed the suitability of several variants of Support Vector Machines (SVM) [10] to topical web page classification tasks, considering them as multi-class problems. We found that supervised approaches

[7] http://www.barnesandnoble.com/

[8] http://www.amazon.com/

[9] In this case, we only consider the tag assignments on books with category data. Considering bookmarks out of those could also reflect on more annotations for one of the user sets, what would be unfair again.

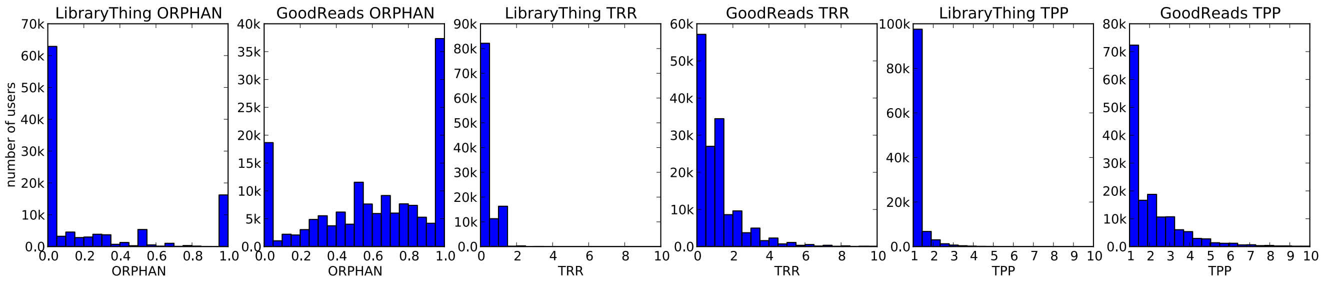

Figure 2: Histogram with distributions of the three studied measures (ORPHAN, TPP and TRR) for the two datasets. X axis represents the quantiles of values, whereas Y axis represents the number of users for each quantile.

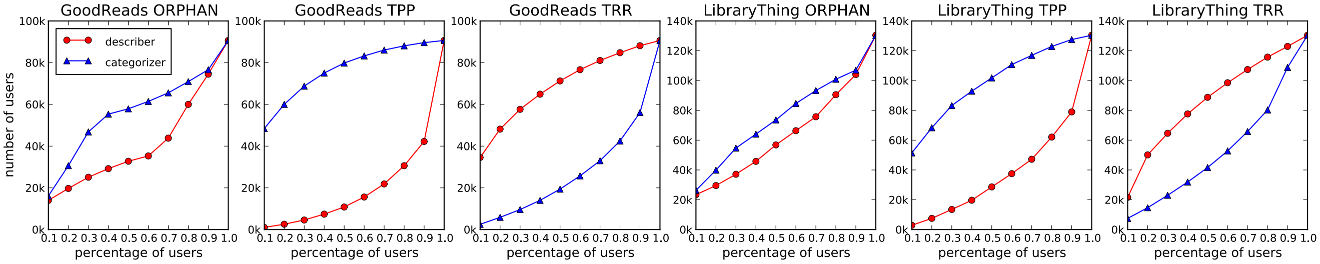

Figure 4: Number of users in each of the subsets. The X axis represents the percents of selected top users, ranging from 10% to 100% of tag assignments, with a step size of 10%, whereas the Y axis represents the number of users. Different user numbers result from the variety of user behavior which are captured by the three measures.

outperform semi-supervised ones, and that considering the task as a single multi-class problem instead of several smaller binary problems performs better. Thus, we use supervised multi-class SVM for our experiments. Even though the traditional SVM approach only works for binary classification tasks, multi-class classification approaches have also been proposed and used in the literature [26] [25]. We use the freely available and well-known “svm-light”[10] in its adapted multi-class version so-called “svm-multiclass”. We set the classifier to work with the linear kernel and the default parameters suggested by the author. The input to the SVM is the set of books in the training set, represented in a Vector Space Model (VSM) where each dimension belongs to a different tag. According to the distribution of these labeled instances in the VSM, a multi-class SVM classifier for k classes defines a model with a set of hyperplanes in the training phase, so that they separate the resources in a category from the rest [2]. The calculation of the hyperplanes is given by Equation 5.

$$\min \left[\frac{1}{2}\sum_{m=1}^{k} ||w_m||^2 + C\sum_{i=1}^{l}\sum_{m\neq y_i}\xi_i^m \right] \qquad (5)$$

Subject to:

$$w_{y_i}\cdot x_i + b_{y_i} \geq w_m\cdot x_i + b_m + 2 - \xi_i^m, \xi_i^m \geq 0$$

where C is the penalty parameter, ξ_i is a stack variable for the i^{th} book, and l is the number of labeled books.

In the test phase, when making predictions for each new resource, the classifier is able to establish a margin for each class. These margins refer to the reliability of the resource to belong to each of the classes. The bigger is the margin, the more likely is the resource to belong to the class. As a result, the class maximizing the margin value will be predicted by the classifier.

We use tags as the input data representing the books to classify using SVM. We use a set of 18,000 books as the training set, whereas the rest (i.e., 9,299 for DDC and 6,861 for LCC) are assigned to the test set. For each of the experiments, we create 6 different runs, choosing different books for the training set on each run. This enables getting more generalistic results, instead of depending on a specific selection of a single training set.

The classifier predicts a category for each book in the test set, according to the side of the hyperplanes they fall into. With classifier’s predictions on all the books in the test set, we compute the accuracy as the percent of correctly classified instances within the test set. As a result, we show the average accuracy of all the runs. The accuracy helps us comparing the extent to which the results of the automated classification resemble to the classification by librarians.

5.2.2 *Descriptiveness of Tags*

To compute the extent to which a set of users is providing descriptive tags, we compare those tags to the descriptive data of books. These descriptive data include the aforementioned synopses, user reviews and editorial reviews. In the first step, we merge all these data in a single text for each book. Accordingly, we get single a text comprising all de-

[10] `http://svmlight.joachims.org`

scriptive data for each book. After this, we compute the frequencies of each term (tf) in the texts, so that we can create a vector for each book, where each of the dimensions in the vectors belong to a term. On the other hand, for each selection of users, we create the vectors of tags for each book, with the annotations of those users. This way, we have the reference descriptive vectors as well as the tag vectors we want to compare to them.

There are several measures that could compute the similarity between a tag vector (T) and a reference vector (R) for a given resource r. They tend to be correlated, though. Regardless of the values given by the measures, we are interested in getting comparable values towards a way to determine whether a tag set resembles to a greater or lesser extent than another set. Thus, as a well-known and robust measure for this, we compute the cosine similarity between the vectors (see Equation 6).

$$\text{similarity}_r = \cos(\theta_r) = \frac{T_r \cdot R_r}{\|T_r\| \|R_r\|} = \sum_{i=1}^{n} \frac{T_{ri} \times R_{ri}}{\sqrt{\sum_{i=1}^{n} (T_{ri})^2} \times \sqrt{\sum_{i=1}^{n} (R_{ri})^2}} \tag{6}$$

The above formula provides the value of similarity between the tag vector and the reference vector of a single book. This value is the cosine of the angle between the two vectors, which could range from 0 to 1, since the term frequencies only consist of positive values. A value of 1 would mean that both vectors are exactly the same, whereas a 0 would mean they don't coincide in neither of the terms, so they are completely different. After getting the similarity value between each pair of vectors, we need to get the overall similarity value between users' tags and descriptions of books. Accordingly, the similarity between the set of n reference vectors, and the set of n tag vectors is computed as the average of similarities between pairs of tag and reference vectors (see Equation 7).

$$\text{similarity} = \frac{1}{n} \sum_{r=1}^{n} \cos(\theta_r) \tag{7}$$

This similarity value shows the extent to which the tags provided by the selected set of users resembles to the reference descriptive data, i.e., how descriptive are the tags by those users. The higher is the similarity value, the more descriptive are the tags provided by the users. The closer it is to 0, the more non-descriptive tags are provided by users.

6. RESULTS

Figure 5 shows the performance of Categorizers (blue line with triangles) and Describers (red line with circles) on the classification task, whereas Figure 6 does the same for the descriptiveness experiments. The results are presented in different graphs separated by datasets, LibraryThing and GoodReads, and by each of the three proposed measures. All of them keep the same scale and ranges for x and y axes, so that it enables comparing the results visually. When analyzing these results, we are especially interested in performance differences between Categorizers and Describers, but also consider other factors, like the degree of improvement between a subset of users, and the whole set. Obviously, both Categorizers and Describers always yield the same performance for 100% sets, as the whole set of users is being considered.

6.1 Categorizers Perform Better on Classification

On the one hand, all three measures get positive results both for the classification and descriptiveness experiments on LibraryThing. The subsets of Categorizers perform better for classification in all cases, whereas Describers outperform for descriptiveness. This means that all three measures provide a good way to discriminate users by behavior. Accordingly, user groups who use tags which are available in the descriptive data perform worse for the classification task than those who do not. Among the compared measures, TPP gets the largest gap for classification, whereas TRR does it for descriptiveness. On the other hand, as regards to GoodReads, results are less consistent. Only TPP provides the results we expected. The others, TRR and ORPHAN, perform well for descriptiveness, but Describers outperform Categorizers for classification. We speculate that the reason for this observation lies in the fact that this social tagging system is suggesting tags to users from their personomy. This encourages users to have a smaller vocabulary, and to reuse their tags frequently, which would effect the overall results. It is quite easier to click on a list of tags than to type them.

6.2 Verbosity vs Diversity

The three measures we have studied in this work rely on two different features to discriminate user behavior: verbosity and diversity. With the better overall performance of the TPP measure as against to the other two, verbosity can be inferred as the optimal feature for discriminating user behavior. In this context, we believe that Categorizers are thinking of shelves when they annotate books with tags, as librarians would do. For instance, a user who thinks of the shelf where she stacks her fictional books seems very likely to solely use the tag *fiction*. We could define these shelf-driven users as non-verbose. A user who adds just one tag has probably thought of the perfect tag that places it in the corresponding shelf. On the other hand, users who provide more detailed annotations rather think of describing the book instead of placing it in a specific shelf. This aspect makes the verbosity feature more powerful than the diversity feature. Thus, we believe that this is the feature that makes TPP so useful as compared to TRR and ORPHAN, because it uniquely relies on users' verbosity.

6.3 The Effect of System Suggestions

We have shown above that, even though all three measures work for LibraryThing, TPP as a measure and verbosity as feature are the only succeeding in a suggestion-biased system like GoodReads. It is worthwhile understanding why diversity is so affected by system's suggestions, though. For instance, a user who has already saved a set of books will face a different annotating task on each system. On LibraryThing, she will have to annotate the book with the tags that come to her mind at the moment of saving it. She will add a few tags if she rather thinks of shelves, and more tags if she wants to describe it, but it is very likely that she will introduce new previously unused tags, because she does not remember her earlier annotations. On GoodReads, however, she will be able to choose and click on a list of tags from her

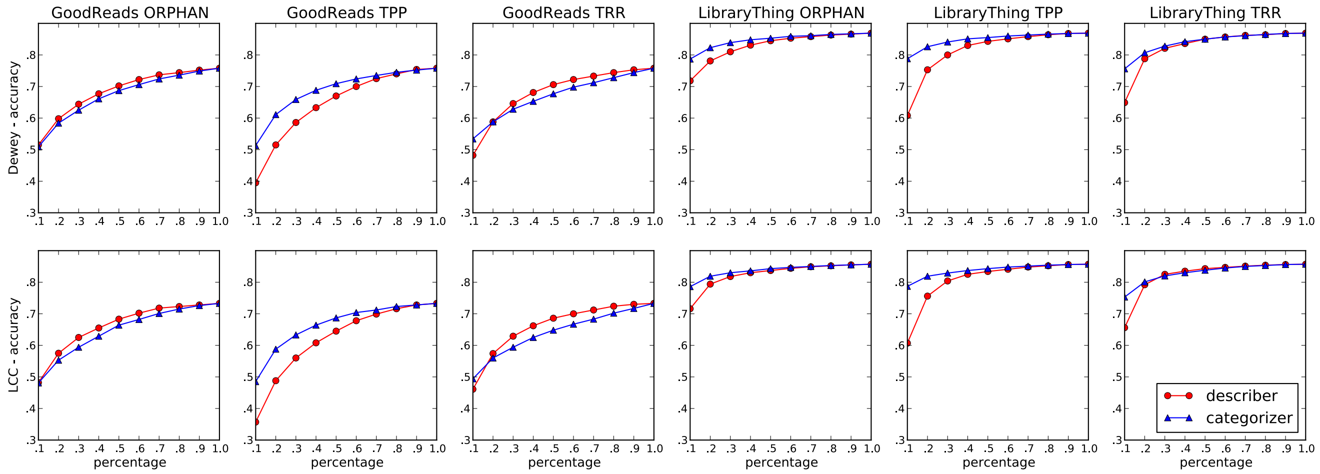

Figure 5: LCC and Dewey Accuracy of LibraryThing and GoodReads. The X axis represents the percents of selected top users, ranging from 10% to 100% and with a step size of 10%, either for Categorizers or Describers, whereas Y axis represents the accuracy. As can be seen *TPP* scores the best accuracy results for Categorizers on both datasets for the two classification schemes. *Orphan* and *TRR* also work for LibraryThing but do not perform for GoodReads.

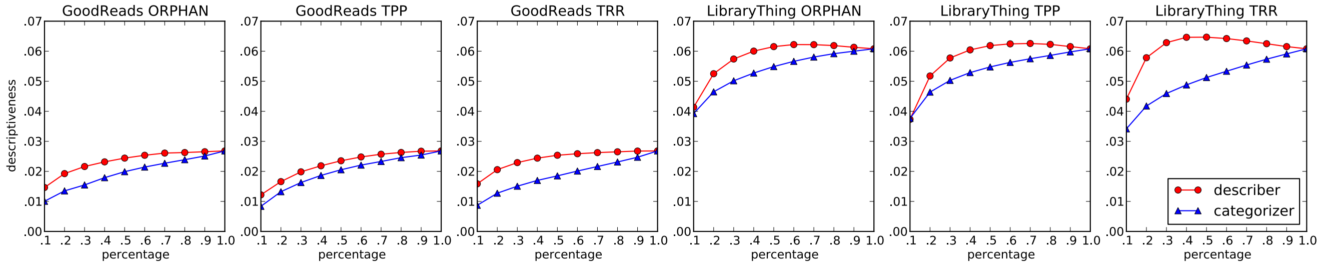

Figure 6: Descriptiveness of LibraryThing and GoodReads. The X axis represents the percents of selected top users, ranging from 10% to 100% and with a step size of 10%, either for Categorizers or Describers, whereas Y axis represents the degree of similarity to descriptive data. When splitting up the folksonomy into Categorizers and Describers, we see that Describers always outperform Categorizers with regard to being similar to the content of metadata.

earlier annotations. She will also choose a few tags if she tends to do it this way, but she will choose many more if she rather describes on her annotations. The main difference from LibraryThing is that she will seldom type new tags like synonyms and other variations because she is looking at the list of available tags. We speculate that this reflects the low effect of suggestions on verbosity, but the high effect on diversity, what makes it dependent on the system.

6.4 Non-descriptive Tags Provide More Accurate Classification

When discriminating user behavior appropriately by using a verbosity-based measure like TPP, we have shown that Categorizers are better suited for the classification task, whereas Describers provide annotations that further resemble to the descriptive data. An interesting deduction from here is that a set of annotations that differs to a greater extent from the descriptive data produces a more accurate classification of the books. From this, we infer that Describers are using more descriptive tags, whereas Categorizers rather use non-descriptive tags. Hence, users who do not think of providing annotations in a similar way to writing reviews rely on non-descriptive tags, yielding a more accurate classification of the books.

Unlike for classification tasks, there are subsets of users who slightly outperform the whole set of users in some cases on descriptiveness. Specifically, this happens with the LibraryThing dataset, and especially when the TRR measure is considered. This means that, in these cases, utmost Categorizers are mainly providing non-descriptive tags.

6.5 Discussion

The use of two different taxonomies for evaluating the classification task make our conclusions more powerful. The results are very similar and comparable from one taxonomy to the other, despite of the considered classification scheme is LCC or DDC. This helps generalize the conclusions and make them non-dependent of the utilized gold standard. Also, working with two social tagging datasets helps understanding how user behavior is affected by interface settings of each system. Hence, a suggestion-biased site like GoodReads has shown to yield very different results from those by LibraryThing. Comparing the classification results by users on these two social tagging systems, we can conclude that tags from LibraryThing outperform tags from GoodReads. In the same manner, tags from LibraryThing seem to saturate the accuracy results when small sets of annotations are being considered. This is not so clear for GoodReads, where larger sets are necessary in order to approach to the best classification accuracy. However, this is likely to happen because of the smaller number of annotations we have for GoodReads, and shouldn't have nothing to do with the behavior of each system's users.

Previous work has shown that the use of social annotations is beneficial in search of an accurate and inexpensive classification of resources [28][3]. These works consider all the users to be equally relevant, though. Going further, our results suggest the existence of users who better fit this kind of tasks. Even though a subset of Categorizers does not outperform the classification accuracy by tags from all the users, the outperformance of Categorizers as compared to Describers should be considered in this context. This evidences that users with non-verbose and non-descriptive behavior provide utmost contributions that give rise to an optimal classification accuracy.

7. CONCLUSIONS

To the best of our knowledge, this paper is the first to report a task, i.e., social classification, in which descriptive tags (produced by Describers) seem *inferior* to non-descriptive tags (produced by Categorizers). Specifically, we have performed both classification and descriptiveness experiments in order to discover how different user behavior effects the performance on certain tasks. Our experiments have been conducted on two social tagging systems that focus on organizing books. For the evaluation process, we have compared users' tags to (1) cataloging data by experts for classification, including the Library of Congress Classification and the Dewey Decimal Classification, and to (2) descriptive book data like synopses and reviews for descriptiveness. While our experiments are limited to the above mentioned data sets and settings, our results warrant future investigations of other datasets, and suggest that the further studies of tagging behavior represent a worthwhile endeavor.

In greater detail, our results show that using verbosity as a feature for discriminating users, Categorizers have shown to be better for the classification task, whereas Describers further resemble to descriptive data. This complements our findings in [28] by further analyzing user-generated annotations insofar as we have found that not all the annotations have the same relevance for the final classification accuracy. Besides this, we have shown that users who do not rely on books' descriptive data provide better classification metadata than those who use descriptive tags. In other words, users who rather annotate with non-descriptive tags more strongly resemble classification performed by expert librarians. This study complements earlier research by identifying relationships between tagging behavior and certain tasks. We found that Categorizers provide more useful tags for the task of classifying them into cataloging schemes. The presented results are relevant for scientists studying social tagging systems and exploring the pragmatics and semantics within these structures as well as designers and developers of social tagging systems who are interested in influencing emerging properties of their systems.

8. FUTURE WORK

We anticipate studying additional means of identifying users who have the potential to enhance the accuracy of classification even further. The exploration of further tagging behavior styles can be a key factor in this context. A potential step stone could be the differentiation between generalists and specialists within social tagging systems. Here specialists could provide better vocabulary which is more focused on the given resource for the classification task as opposed to generalists who annotate resources with general tags.

9. ACKNOWLEDGMENTS

The research presented in this work is in part funded by the Know-Center, the FWF Austrian Science Fund Grant P20269, the European Commission as part of the FP7 Marie Curie IAPP project TEAM (grant no. 251514), the Regional Government of Madrid under the Research Network MA2VICMR (S-2009/TIC-1542), the Regional Ministry of Education of the Community of Madrid, by the Spanish Ministry of Science and the Innovation project Holopedia (TIN2010-21128-C02-01).

10. REFERENCES

[1] E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, New York, NY, USA, 2008. ACM.

[2] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, 2002.

[3] D. Godoy and A. Amandi. Exploiting the social capital of folksonomies for web page classification. In *Software Services for E-World*, volume 341 of *IFIP Advances in Information and Communication Technology*, pages 151–160. Springer, 2010.

[4] S. Golder and B. Huberman. The structure of collaborative tagging systems. *Arxiv preprint cs/0508082*, 32(2):198–208, 2005.

[5] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (I). *D-Lib Magazine*, 11(4):1082–9873, 2005.

[6] M. Heckner, M. Heilemann, and C. Wolff. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *International AAAI Conference on Weblogs and Social Media (ICWSM)*, San Jose, CA, USA, May 2009.

[7] M. Heckner, T. Neubauer, and C. Wolff. Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types. *Conference on Information and Knowledge Management*, 2008.

[8] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *First ACM International Conference on Web Search and Data Mining (WSDM'08)*, February 2008.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Folkrank: A ranking algorithm for folksonomies. In *Proc. FGIR 2006*, pages 111–114, 2006.

[10] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, pages 137–142, Berlin, 1998. Springer.

[11] C. Körner. Understanding the motivation behind tagging. ACM Student Research Competition - Hypertext'09, 2009.

[12] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *International World Wide Web Conference*, pages 521–530, 2010.

[13] C. Körner, R. Kern, H.-P. Grahsl, and M. Strohmaier. Of categorizers and Describers: an evaluation of quantitative measures for tagging motivation. In *Conference on Hypertext and Hypermedia*, pages 157–166, 2010.

[14] M. Lipczak and E. Milios. The impact of resource title on tags in collaborative tagging systems. *Conference on Hypertext and Hypermedia*, pages 179–188, 2010.

[15] C. Lu, J.-r. Park, and X. Hu. User tags versus expert-assigned subject terms: A comparison of librarything tags and library of congress subject headings. *Journal of Information Science*, 36(6):763–779, 2010.

[16] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[17] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *International Semantic Web Conference*, LNCS, pages 522–536. Springer, 2005.

[18] M. G. Noll and C. Meinel. Exploring social annotations for web document classification. *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08*, page 2315, 2008.

[19] M. G. Noll and C. Meinel. The metadata triumvirate: Social annotations, anchor texts and search queries. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, pages 640–647, Washington, DC, USA, 2008. IEEE Computer Society.

[20] E. Rader and R. Wash. Influences on tag choices in del.icio.us. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer Supported Cooperative Work*, pages 239–248, New York, NY, USA, 2008. ACM.

[21] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*, November 2008.

[22] G. Smith. *Tagging: people-powered metadata for the social web*. New Riders, Berkeley, Calif. :, 2008.

[23] M. Strohmaier, C. Körner, and R. Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. In *International AAAI Conference on Weblogs and Social Media (ICWSM2010)*, Washington, DC, USA, May 2010.

[24] R. Wash and E. Rader. Public bookmarks and private benefits: An analysis of incentives in social computing. In *ASIS&T Annual Meeting*. Citeseer, 2007.

[25] C. wei Hsu and C. jen Lin. A comparison of methods for multi-class support vector machines. 2007.

[26] J. Weston and C. Watkins. Multi-class support vector machines. In *Proceedings of the 1999 European Symposium on Artificial Neural Networks*, 1999.

[27] A. Zubiaga, V. Fresno, and R. Martínez. Is unlabeled data suitable for multiclass svm-based web page classification? In *SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 28–36, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[28] A. Zubiaga, R. Martínez, and V. Fresno. Getting the most out of social annotations for web page classification. *Document Engineering*, pages 74–83, 2009.