Aspects of User Motivation in Social Tagging Systems

Dipl.-Ing. Christian Körner, Bakk.rer.soc.oec.

DISSERTATION

zur Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften
der Studienrichtung
Informatik
erreicht an der
Technischen Universität Graz

Univ.-Doz. Dipl.-Ing. Dr.techn. Markus Strohmaier
Institut für Wissensmanagement
Technische Universität Graz

Graz 2012

For Christine

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am ..

(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

date ..

(signature)

Abstract

Tagging describes the process of annotating resources with terms - so called “tags” - in order to enable better organization, retrieval and sharing of information. Today, a large number of diverse applications exist that allow users to annotate digital information from their libraries on the web. These systems are referred to as social tagging systems. Delicious, for example, enables users to organize their bookmarks and tag them accordingly. On Flickr, users can tag pictures and on YouTube content creators have the possibility to assign tags to videos in order to facilitate retrieval. While these systems have increasingly become a subject of research, little is known about how and why these systems are used and how motivation and the resulting behavior of individual users yield emergent system properties. This thesis introduces a new distinction between two kinds of tagging motivation - categorization and description. While recent studies on the motivation behind tagging exist, they are based on questionnaires or the judgements of experts. An automated way of surveying tagging motivation in these social structures is not yet available. This work introduces a quantitative analysis of tagging motivation by evaluating statistical properties found in a user’s tag vocabulary. For this purpose a number of measures to facilitate the distinction of these user types are presented and quantitatively as well as qualitatively evaluated. Ultimately, it is shown that certain aspects of tagging motivation can be approximated by simple statistical measures. Further work investigates how data provided by the two different kinds of users - categorizers and describers - affects different tasks of knowledge extraction from social tagging systems, in particular capturing semantic relations and automated classification. The results of the experiments conducted show that data of categorizers is better suited for social classification tasks than their counterparts who are describing documents from their collection. Furthermore, users driven by description

provide better input for the emergence of semantics in a social tagging system. Based on these results first evidence for a causal link between the pragmatics and the semantics of such systems is found. This indicates that user behavior affects the structure of tagging systems - information that can be used for system designers and architects of these platforms. As an outlook, an additional evaluation is presented that can serve as a step stone for a future expert identification task. The contribution of this work is the introduction of and differentiation between two types of tagging motivation along with the presentation of associated measures to distinguish them. The application of these measures to real world datasets shows that tagging motivation varies within and across social tagging systems. Further analyses investigate the effects user groups exhibiting certain behavior have on knowledge extraction tasks in tagging systems. This work is relevant for researchers and system designers interested in user motivation in social tagging systems, and the analysis of resulting consequences these characteristics have on the contained social structures.

Zusammenfassung

Tagging bezeichnet das Annotieren von digitalen Ressourcen mit Schlagworten - so genannten "Tags" - mit dem Zweck Information besser zu organisieren, leichter wiederfindbar zu machen und deren gemeinsame Nutzung zu ermöglichen. Gegenwärtig existiert im Web eine große Anzahl von Applikationen die es Benutzern erlauben Informationen zu annotieren. Diese Systeme werden soziale Taggingsysteme genannt. Delicious beispielsweise erlaubt Benutzern ihre Lesezeichen mit Hilfe von Tags zu organisieren. In Flickr können Benutzer Bilder verschlagworten und YouTube ermöglicht die Vergabe von Tags um Videos innerhalb des Systems leichter auffindbar zu machen. Obwohl diese Systeme in den letzten Jahren zunehmend in den Fokus der Forschung gerückt sind, ist noch immer wenig über die Verwendung von Tags innerhalb dieser Plattformen und die damit einhergehenden Absichten der Benutzer bekannt. In weiterer Folge gibt es auch keine Studien darüber wie sich die Motivation von Benutzern und das daraus resultierende Verhalten in den Eigenschaften eines solchen Systems widerspiegeln. Die vorliegende Arbeit führt die Unterscheidung zweier neuer Arten von Taggingmotivation ein - Beschreibung und Kategorisierung. Bisher verfügbare Arbeiten, die sich mit der Analyse von Motivation in sozialen Taggingsystemen beschäftigen, basieren entweder auf der Einschätzung von Experten oder der Auswertung von Fragebögen. Bis heute existiert keine automatisierte Untersuchung von Taggingmotivation in diesen Systemen. Diese Dissertation stellt eine quantitative Analyse von Benutzermotivation vor, bei der statistische Eigenschaften des Tagvokabulars eines Benutzers untersucht werden. Für die Unterscheidung der zwei Arten von Taggingmotivation werden eine Reihe von Methoden eingeführt und sowohl qualitativ als auch quantitativ evaluiert. Als Resultat dieser Untersuchungen wird die Messung von Taggingmotivation mithilfe einfacher statistischer Größen ermöglicht. In zusätzlichen Experimenten

wird analysiert, wie sich Daten der zwei Benutzergruppen unterschiedlich auf verschiedene Methoden der Wissenserschließung in Taggingsystemen auswirken. Besonderes Augenmerk wird hierbei auf die automatische Klassifikation sowie das Erfassen von Semantik gelegt. Die Resultate der Experimente zeigen, dass Kategorisierer besser für soziale Klassifikationszwecke geeignet sind, während Beschreiber besser zu der in Taggingsystemen auftretenden Semantik beitragen. Diese Ergebnisse zeigen einen Zusammenhang zwischen der Verwendung dieser Systeme und der in ihnen vorkommenden Semantik. Dies lässt darauf schließen, dass sich das Verhalten von Benutzern auf die Struktur dieser Systeme auswirkt - Information die besonders für Designer und Architekten dieser Plattformen von Bedeutung ist. Der wissenschaftliche Beitrag der vorliegenden Arbeit liegt in der Einführung und Unterscheidung zweier Arten von Taggingmotivation und der damit verbundenen Methoden um zwischen ihnen zu differenzieren. Mithilfe einer Auswertung auf mehreren Taggingdatensätzen wird gezeigt, dass Taggingmotivation sowohl innerhalb einzelner als auch zwischen unterschiedlichen Plattformen variiert. Des Weiteren wird der Einfluss der einzelnen Taggingmotivationsgruppen auf Verfahren der Wissenserschließung analysiert. Diese Arbeit ist relevant für Wissenschafter und Systemdesigner die an Benutzermotivation in sozialen Taggingsystemen und den daraus resultierenden Auswirkungen interessiert sind.

Acknowledgements

I would like to sincerely thank my PhD adviser Markus Strohmaier for his input and guidance in the past years. Without his excellent feedback this would not have been possible.

Further, I am very grateful to Dominik Benz from KDE Kassel and Arkaitz Zubiaga from Universidad Nacional de Educación a Distancia Madrid for excellent discussions and fruitful cooperations. It is great to to find fellow researchers who are not only brilliant in the things they do but grow friends with whom one can go for drinks after a long day of work.

I also want to thank the Knowledge and Data Engineering Group - University of Kassel for the warm welcome and the brilliant collaboration during my time in Kassel. Specifically I like to thank Andreas Hotho and Gerd Stumme for giving insights and inspiration.

Further I am grateful to Kris Jack and all employees of Mendeley LTD for showing me the London start-up scene and providing me with a high productive workplace environment and countless interesting nights during my six month stay in London.

I would also express my gratitude to the members of the Knowledge Management Institute and the Know Center at Graz University of Technology. Thank you for giving me a home in the past years and helping me become the researcher I currently am. Especially I would like to thank Roman Kern, Denis Helic, Claudia Wagner, Christoph Trattner, Karin Schöfegger, Mark Kröll, Hans-Peter Grahsl and Philipp Singer for their help and constructive feedback on the dissertation at hand.

Last but not least I want to thank my family for their support and help during the years of my PhD studies. Christine, thank you for your kindness, patience and your ability to make me smile. You are my constant.

The presented research was funded by the FWF Austrian Science Grant P20269 *TransAgere* - "Agent-Oriented Engineering of Social Software", the European FP7 Project *TEAM* - "Transferring Knowledge in Academic Knowledge Management" and the FWF Austrian Science Grant I677 *PoSTs* - "Pragmatics and Semantics in Social Tagging Systems".

Contents

1. Introduction

1.1. Motivation

The advent of the *Web 2.0* marks the beginning of easy and efficient content creation and consumption on the World Wide Web. For the first time, users who are not technically knowledgeable are able to upload videos, write blog entries and annotate resources in an intuitive and easy manner without deep knowledge of scripting languages such as JavaScript or understanding of markup languages like HTML. Organization of digital resources is improved by new applications such as *social tagging systems* that enable users to catalog, annotate, browse and re-find resources of their own and other users' libraries by using freely chosen terms - so called "tags".

These platforms can be seen as a low cost alternative to traditional classification systems such as the Dewey Decimal Classification System (short DDC) [Dewey, 1876] and the Library of Congress Classification system[1] (short LCC). But while these traditional classification methods require extensive training for users to be able to catalog or retrieve documents, tagging requires hardly any training and is cognitively relatively inexpensive. Moreover, it enables the distribution of the annotation process on a multitude of users and therefore allows an open range of free form metadata to be associated to digital resources.

However, due to the nature of this annotation process a number of challenges emerges. The uncontrolled and free-form vocabulary in addition to the dynamics found within these social platforms require special methods to generate meaning for tags. Further, it is difficult to elicit structure

[1] http://www.loc.gov/catdir/cpso/lcc.html

from the systems that include tagging because user behavior and the underlying motivation influence the results.

While in the past a lot of research has been conducted on the fabric and emergent semantic structures of folksonomic systems, the reasons how and why users are tagging still remain largely elusive. Even though empirical studies on motivation in these social systems already exist, to the best of our knowledge there is no automated way of surveying tagging motivation. Neither do we know how tagging motivation observed by groups of users effects the resulting social tagging system as a whole nor how data provided from user groups driven by diverse motivations perform on tasks like semantic extraction and automatic classification.

This dissertation introduces a quantitative analysis of user motivation in social tagging systems and evaluates the resulting influence user groups of different motivations on folksonomies which represent the tripartite structure of users, tags and resources (for a more detailed definition see Section 2.1.1).

The goal of this first chapter is to provide an introduction into the concept of tagging, its properties and associated potentials as well as limitations. Further, it gives an selective overview of current social tagging systems. Subsequently, the research questions of this dissertation are elaborated and the structure of this dissertation outlined.

1.2. Social tagging

Tagging describes the process of assigning tags to any kind of digital resources such as URLs, photos, products, videos and other digital information in social systems. *Social tagging* is the collaborative annotation effort made by users within social tagging systems.

Tags are non-hierarchical free-form terms chosen by the users based on their own preferences and decisions. Usually, users are allowed to use any sequence of characters apart from those which are defined as tag delimiters by the designers of the system (e.g. the “space” character in Delicious and BibSonomy). While there exists a couple of social tagging

systems that enforce certain conventions for the tags used within them (see Section 1.2.2, for examples) the vast majority let users freely decide on chosen tags. In comparison to traditional classification systems that use predefined categories or hierarchical structures, tagging enables the users to assign tags in a free and unbound manner.

Tagging is often referred to as a "*bottom up*" approach in which users are not restricted to predefined terms or categories ([Hammond et al., 2005]). This is in contrast to "*top down*" processes such as traditional classification systems (e.g. the previously mentioned the Library of Congress Classification system) in which users need to have extensive training in order to be knowledgeable of the underlying structure and methodology for finding and cataloging resources.

1.2.1. Example

In the following an example for the process of tagging is given. The URL `http://www.reddit.com` serves as the resource that is going to be annotated on the Delicious system.

Figure 1.1 shows the "Save Bookmark" dialog found in Delicious. This form is displayed when a user decides to bookmark a web site manually or via the bookmarklet[2]. The first text field (1) contains the title of the selected resource. In the second field (2), users are able to type in tags that are optionally auto-completed with suggestions already found in the user's tag vocabulary. Below the tag field the system recommends further tags to support the user (3). These tags are either already contained in the user's collection, were applied to the resource by others in the system or supplied by a tag recommendation algorithm. Upon selecting a recommended tag, it is automatically added to the manually entered tags in the text field. By clicking "Save" the URL and the chosen tags are added to the user's library.

[2] A bookmarklet is a link found in a browser's bookmarks that contains JavaScript and provides one-click functionality to e.g. store photos online.

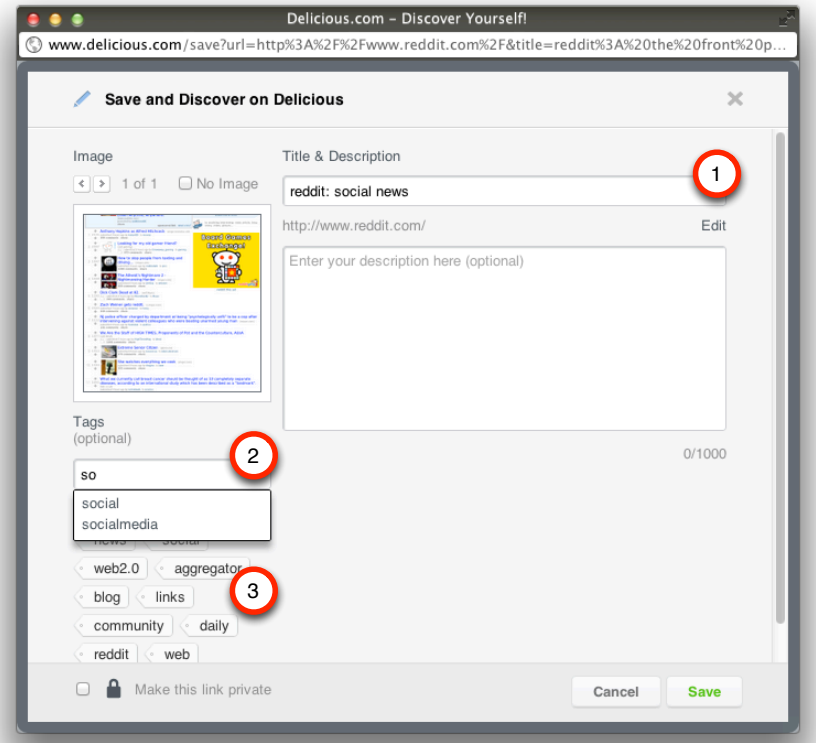


Figure 1.1.: "Save Bookmark" dialog of Delicious displaying the annotation of the website http://www.reddit.com with Tags

1.2.2. Social tagging systems

In the past years a lot of popular online platforms emerged that include tagging as a primary feature or added it to allow users a variety of tasks such as the annotation or description of digital resources and the subsequent retrieval thereof. This section aims to give a short overview of the types of tagging systems that currently exist and briefly examines their properties.

Online bookmarking tools enable their users to store URLs of pages they want to share or retrieve later. The most prominent example in the con-

text of social tagging is *Delicious*[3] - a social bookmarking site. Figure 1.2 shows an excerpt of resources tagged by different users in this system. Below the resources the associated list of tags is presented. On the right side the number of users who have a resource in their libraries is displayed.

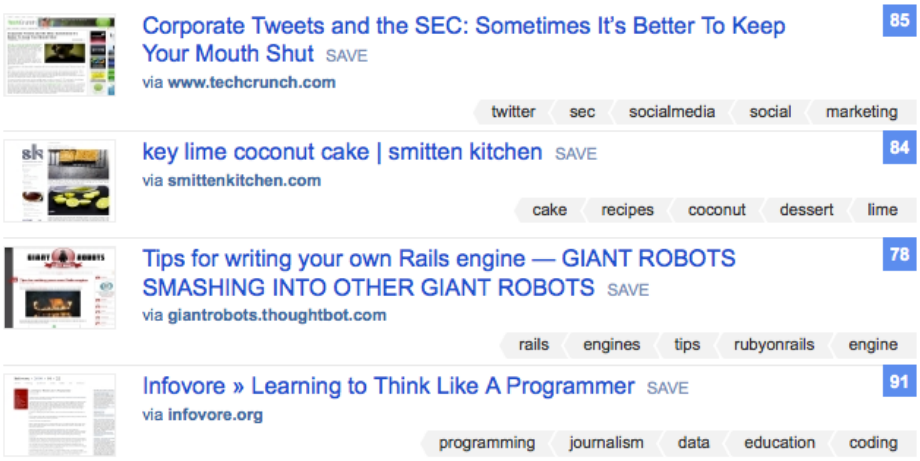


Figure 1.2.: A list of resources annotated by users of the Delicious system

A popular visualization technique is tag clouds that illustrates the vocabulary of a resource or a collection of documents by displaying a set of tags. The importance of each tag - usually defined by the number of occurrences - is reflected in its font size. The larger a tag is displayed the more important it is in the presented context. Figure 1.3 shows an example of a tag cloud found in the author's Delicious library[4].

In *Photo Sharing Systems* like Flickr[5], Picasa[6] and 500px[7] users are able to upload, organize and share their photos. Within these platforms tags are used to describe the contents of pictures (e.g. "`dog`", "`high definition`" and "`family`)", express feelings or opinions (e.g. "`awesome`", "`Top100`") or to catalog photos into events ("`holidays`", "`thanksgiving`" etc.).

[3] http://www.delicious.com
[4] https://delicious.com/chriskoerner/tags?sort=alpha
[5] http://www.flickr.com
[6] http://picasaweb.google.com
[7] http://www.500px.com

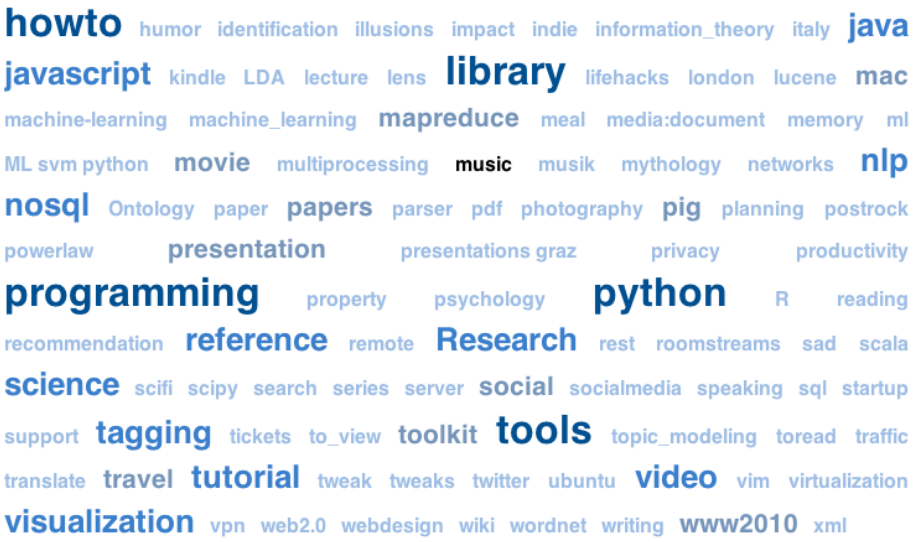


Figure 1.3.: Excerpt of a user's vocabulary visualized via a tag cloud in Delicious

Another use case for tagging is *Bibliographic Management*. Tools for bibliographic management support users who are writing scientific article and theses. Online systems like BibSonomy[8], CiteULike[9] and Mendeley[10] feature either the storage of bibliography references or the document including the associated metadata and provide tagging as an additional aspect for annotation and categorization. Tags either give insights into the topic (e.g. "`information_retrieval`") or shed light on the organization of users ("`to_read`", "`my_own`").

Apart from the presented systems there exists a large number of other online platforms that include tagging as a supplementary feature. Examples for such platforms are:

- In the **YouTube**[11] system content creators are able to add tags as metadata to their uploaded videos to make them easier to find and to support navigation.

[8] http://www.bibsonomy.org
[9] http://www.citeulike.com
[10] http://www.mendeley.com
[11] http://www.youtube.com

- **Amazon**[12] allows customers to tag products to introduce an additional facility for exploration. In addition customers are able to navigate through tags that are related to the currently inspected tag.
- Users of the micro-blogging platform **Twitter**[13] use *hashtags* (e.g. "`#wikipedia`") in their streams to signify that the tweet belongs to a specific stream, topic or event.

Next to traditional social tagging systems exist a couple of derivatives that restrict the free form of annotation or propose rules to enable consistent vocabulary throughout the platform: *Faviki*[14] for instance requires tags to be concepts from Wikipedia[15] to minimize the ambiguity. Due to this disambiguation step that is supported by a recommendation mechanism, annotation in this system takes significantly longer than in systems that do not have this constriction. Another example is the question answering system *StackOverflow*[16] in which users can apply tags to their threads to assign them to categories. While it is allowed to choose tags freely, the developers of the system recommend the usage of normalized vocabulary in order to make questions easier findable by other users who can then in turn provide answers.

Since the availability of datasets for research is still quite scarce and data gathering is error prone, time consuming and technically challenging a list of datasets which are available for academic research was collected by us in [Körner and Strohmaier, 2010].

1.2.3. Opportunities and challenges of social tagging

Due to the addition of free-form annotations as meta data, social tagging systems face a number of potentials and challenges (cf. [Mathes, 2004] and [Quintarelli, 2005]). Some of the potential advantages found in these platforms are:

[12] `http://www.amazon.com`
[13] `http://www.twitter.com`
[14] `http://www.faviki.com`
[15] `http://www.wikipedia.org`
[16] `http://www.stackoverflow.com`

- *Cognitive costs* - The process of tagging is quickly explained, very easy to grasp and cognitively relatively inexpensive.
- *Serendipity* - The diversity of tags enables browsing and finding resources by exploring tags and their relations.
- *Personal vocabulary* - Users can create their own vocabulary and are not limited to taxonomic properties of traditional categorization systems.
- *No authority* - No central authority enforces rules for the use of tags in such systems.
- *Folksonomies are inclusive* - Each tag assignment is reflected immediately in the tagging vocabulary of the folksonomy without the omission of any tags. In combination with the power law properties that emerge in these systems the discovery of long tail concepts is possible.
- *Multiple Viewpoints* - When a resource is tagged by different users, tags provide insight into other views from other people.

The ease of applying tags to resources comes with a number of challenges. Due to the free and unbound vocabulary that is used for the tag application a couple of challenges emerge in these systems:

- *Ambiguity of tags* - Since there are no restrictions or rules for the usage of tags users introduce a variety of synonyms and homonyms into the system.
- *Language independence* - Users are allowed to express their descriptions and annotations in their own or a language of their choice leading to the availability of multiple translations of the same word.
- *Lack of ruleset for formal writing of tags* - There exist no rules how to write compound tags. "`SemanticWeb`", "`Semantic_Web`" and "`Semantic-Web`" are all valid tags that can be found in tagging systems.
- *No relationship between tags* - In contrast to classification systems like the Dewey Classification folksonomies lack relations between

categories (e.g. a hierarchical structure).

- *Levels of abstractness* - A direct result of the previous problem is the absence of indications how specific or general a tag is (see Section 3.9 for possibilities to tackle this challenge).
- *Spelling errors* - Typing errors can leave tags in the folksonomy that are "orphans" and of little or no use for the users.
- *Scaling* - The scaling of tags is a problem when retrieving documents from a large collection of resources.

Tackling all the problems is beyond the scope of this thesis, but we will show that the differentiation of users based on different types of tagging motivation has influences on the performance of existing mechanisms for extracting semantics and social classification.

1.3. Problem statement and research questions

This work aims to investigate tagging behavior and the underlying motivation for tagging in social tagging systems:

> The exploration of types of tagging motivation, possibilities to measure them and the analysis of their influence on social tagging systems are the main objectives of this dissertation. Of special interest is the impact different user motivations have on tasks such as the capturing of semantics or social classification within these systems.

Given a social tagging system comprising users, tags and resources, we are interested in finding answers for the following research questions:

1.3.1. RQ1 - What kinds of tagging motivation can be identified in social tagging systems?

When examining social tagging systems, the analysis of different users and their libraries is possible. The objective of this research question is

the identification of different forms of tagging motivation found in folksonomies. This helps not only to shed light on the reasons why users are using tags on these platforms but also enables their designers and developers to align their platforms and e.g. support users accordingly.

The paper "Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems" [Strohmaier et al., 2010] (Section 3.4) addresses this question by introducing two different types of tagging motivation. *Describers* are users driven by the motivation of annotating resources in a descriptive and verbose manner whereas *categorizers* use tags as replacement for categories and try to establish a consistent vocabulary.

1.3.2. RQ2 - Is it possible to measure tagging motivation automatically?

Based on the identification of different types of tagging motivation, the next step is to examine if it is possible to evaluate them with the help of automatic mechanisms. An advantage of such an analysis is that it does not necessitate time-expensive empirical studies which need human subjects or the laborious evaluation of questionnaires.

The papers that deal with this question are "Exploring the Influence of Tagging Motivation on Tagging Behavior" [Kern et al., 2010] (Section 3.5) and "Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation" [Körner et al., 2010b] (Section 3.6). The papers present a range of different statistical measures for the differentiation of the two types of tagging motivation. Further, the measures are quantitatively and qualitatively evaluated to find those that capture the differentiation best.

1.3.3. RQ3 - How does tagging motivation vary within and across systems?

Equipped with the knowledge of how to measure tagging motivation from user behavior, it is interesting to examine how different motivation types

occur in single social tagging systems as well as how they vary across multiple social tagging systems. This way it is possible to investigate what types of tagging motivation occur on a wide range of systems and to hypothesize how tagging motivation is influenced by properties (such as tag recommenders) of these applications.

"Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems" [Strohmaier et al., 2010] (Section 3.4) as well as "Exploring the Influence of Tagging Motivation on Tagging Behavior" [Kern et al., 2010] (Section 3.5) are the papers that cover this research question. In these papers, the previously introduced measures are applied to a range of different tagging datasets such as Delicious, Diigo[17], BibSonomy and Flickr in order to evaluate tagging motivation in these systems.

1.3.4. RQ4 - What effects does tagging motivation have on knowledge extraction tasks like social classification and semantic extraction?

Our knowledge about different types of tagging motivation and measures to detect them enables the investigation of what tasks the different user groups are better suited for. The objective of this research question is to investigate if certain forms of user motivation generate better prerequisites for extracting semantic relations or provide data that performs better for classification. This information is of special interest for researchers and architects of folksonomic systems and would allow the computation of these automated tasks on a subset of users instead of the complete data.

The papers investigating this research question are: "Stop Thinking, Start Tagging: Tag Semantics Emerge from Collaborative Verbosity" [Körner et al., 2010a] (Section 3.7) and "Tags vs Shelves - From Social Tagging to Social Classification" [Zubiaga et al., 2011] (Section 3.8). The first paper studies the influence of tagging motivation on the emergence of semantics in folksonomic systems. The second paper analyzes how user

[17] http://www.diigo.com

motivation of social tagging systems affects the task of social classification.

1.4. Organization of this thesis

This cumulative thesis consists of the following papers:

- Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems [Strohmaier et al., 2010]
- Exploring the Influence of Tagging Motivation on Tagging Behavior [Kern et al., 2010]
- Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation [Körner et al., 2010b]
- Stop Thinking, Start Tagging: Tag Semantics Emerge from Collaborative Verbosity [Körner et al., 2010a]
- Tags vs Shelves - From Social Tagging to Social Classification [Zubiaga et al., 2011]
- One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata [Benz et al., 2011]

A full list of paper co-authored can be found in Appendix A.

The thesis at hand is organized as follows: The current Chapter (Chapter 1) gives an introduction, states the problem setting and the research questions of this dissertation. Chapter 2 presents the used terminology and an overview of related work. This is followed by Chapter 3 which contains the papers of this cumulative dissertation. Finally, Chapter 4 elaborates results and contributions as well as answers to the research questions and points to future work.

1.4.1. Graphical illustration of the organization of the thesis

Figure 1.4 depicts a graphical illustration of the content found in this dissertation. It shows how research questions relate to the topics addressed

in this work and highlights how individual papers relate to these topics. The paper "One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata" is a step towards future work and therefore does not directly relate to the research questions which are introduced in Section 1.3.

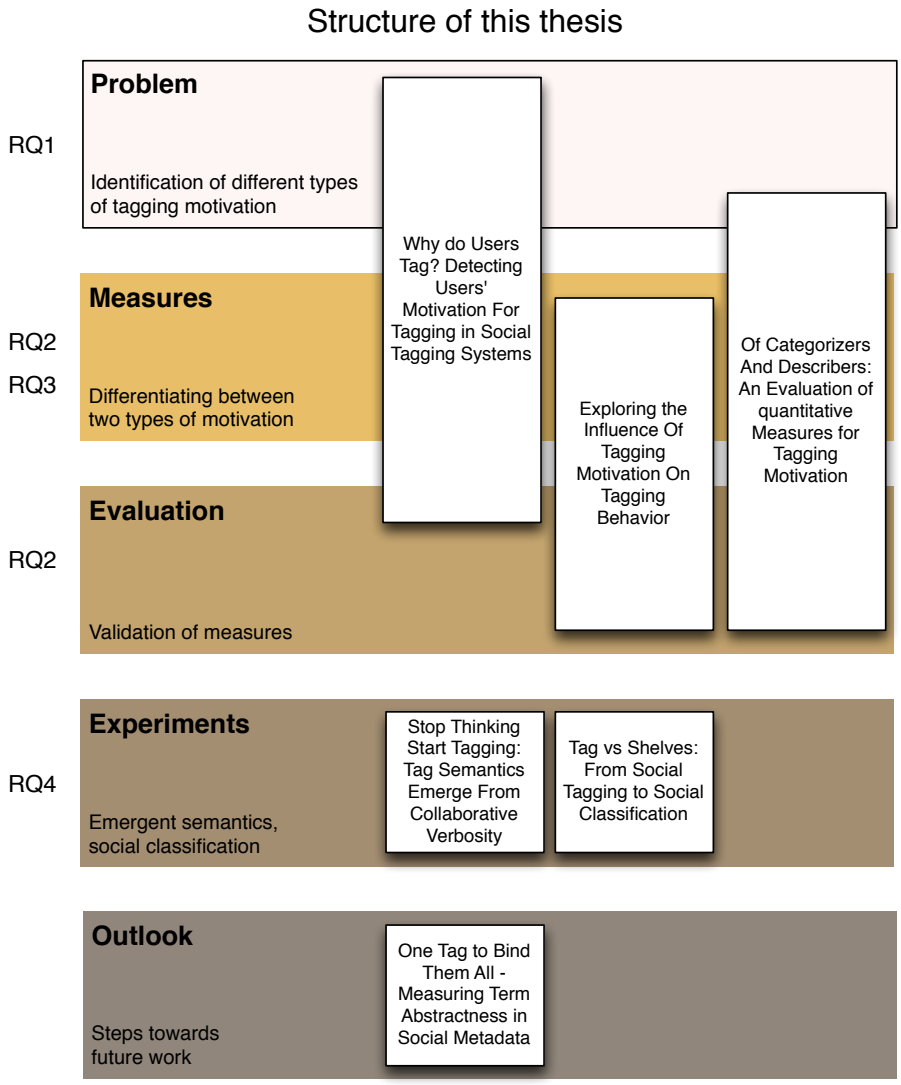


Figure 1.4.: Relationship of research questions, topics and associated papers

2. Terminology and related work

2.1. Terminology

In this section the terminology used in the papers comprising this cumulative dissertation is briefly outlined.

2.1.1. Folksonomy

The term *folksonomy* was coined by [Wal, 2007] and combines the two terms "*folk*" and "*taxonomy*" - referring to the consensus that emerges when users label resources with tags in social tagging systems.

In general, folksonomies are represented as tripartite graphs with hyper edges. The three types of disjoint sets found in these structures are:

- Users - The set of *users* $u \in U$
- Tags - The set of *tags* $t \in T$
- Resources - The set of *resources* $r \in R$

Subsequently a folksonomy is defined as the complete set of annotations $F \subseteq U \times T \times R$. A *personomy* represents the reduction of the folksonomy F to a particular user $u \in U$ with all her tags and resources (cf. [Hotho et al., 2006]). The triple consisting of a single user $u \in U$, a tag $t \in T$ and a resource $r \in R$ is called *tag assignment* (in short TAS - $tas \in TAS$). Figure 2.1 shows an exemplary small folksonomy with the three sets of users, tags and resources. A link between a user and a tag indicates the usage of the given tag by the particular user. A connection between a tag and a resource signifies the annotation of a resource with the corresponding tag. An example for a tag assignment would be

the triple of user C, the tag "`wiki`" and the resource found on the lower right.

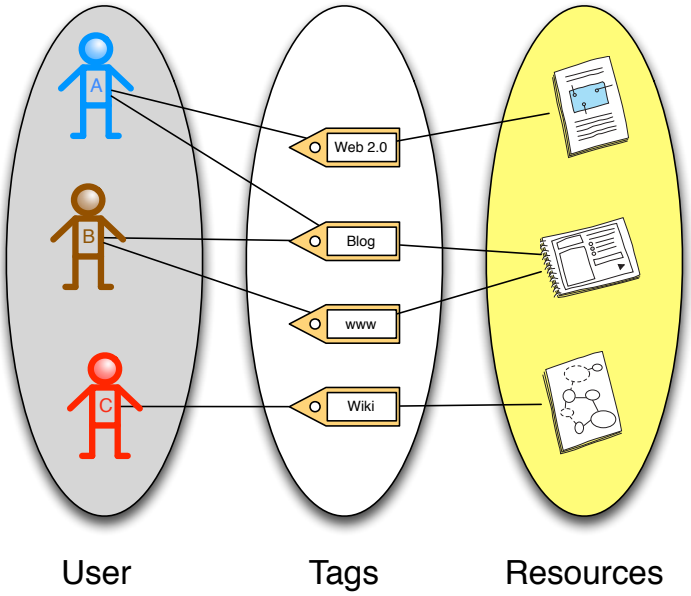


Figure 2.1.: Exemplary folksonomy with the three sets of users, tags and resources

It is important to note that the term "folksonomy" is used ambiguously throughout the scientific community. Other work uses the term folksonomy in a slightly different way (cf. [Plangprasopchok and Lerman, 2008] and [Helic et al., 2011]). Here, a folksonomy is seen as the hierarchical classification that emerges from the tripartite structure of users, tags and resources.

2.1.2. Broad and narrow folksonomies

According to [Wal, 2005] social tagging systems can be differentiated in two classes - broad and narrow folksonomies. In *broad folksonomies* all users are able to annotate resources in a free manner. Resources can occur in multiple user collections and are therefore often tagged by a multitude of people. An example for such a broad folksonomy is the Delicious sys-

tem. Due to the multi-user nature of this system popular resources are often seen with synonym or redundant tags as metadata.

In *narrow folksonomies* only one user - typically the content creator or uploader - can apply tags to resources. This can either be because the system enforces it or because the users wishes his library and included tags to be private. To give some examples: In YouTube video uploaders are able to specify tags for their resources to make them easier findable by others. However, other users are not able to tag videos. *Mendeley* is another example for a narrow folksonomy which allows users to catalog their private copies of scientific papers with tags. These tags are not directly visible for others and serve therefor primarily a personal organizational purpose.

Another definition for this categorization scheme of social tagging systems is given by [Smith, 2008] who defines narrow folksonomies as *simple* and broad folksonomies as *collaborative* tagging systems.

2.1.3. Tagging motivation

This work defines *tagging motivation* as the reasons and goals users have when annotating resources in social tagging systems. These reasons depend both on the audience (e.g. annotation for the user herself or others) and on the task that should be accomplished by the utilization of tags (e.g. retrieval, browsing, replacement for folders). A user is motivated to behavior in a particular way. We can observe and analyse this behavior.

Examples for such types of motivation are "personal information management" versus "resource sharing" as introduced by [Heckner et al., 2009] and "enjoyment", "commitment", "self reputation" etc. as shown by [Nov et al., 2009]. Further details on these and others categories of tagging motivation are found in Section 2.2.

2.2. Related work

For this dissertation the following research topics are relevant: the analysis of folksonomies, the studies of tagging behavior in social tagging systems, models as a tool to investigate folksonomic systems and cognitive processes during the course of tagging. This chapter is intended to give a high-level, incomplete introductory overview of these areas. For further details see the corresponding sections in the papers included in this dissertation.

2.2.1. Folksonomy analysis

Folksonomy analysis deals with the study and evaluation of social tagging systems and examines the contained structure including the relations of the comprising users, tags and resources.

[Hammond et al., 2005] were the first to conduct early analyses of folksonomic systems. The authors examined nine different social bookmarking platforms such as CiteULike, Connotea[18], Delicious, Flickr and others. In this work two dimensions of tagging methodology are differentiated: *tag creation* and *tag usage*. Figure 2.2 shows the two dimensions and their interrelation. The figure indicates that Flickr users tend to use their tags for their own purposes whereas in the other systems users tag content which was not created by themselves.

The best known and most influential paper on surveying social tagging systems is [Golder and Huberman, 2006] in which another early analysis of folksonomies is conducted. The authors investigate the structure of collaborative tagging systems and found regularities in user activity, tag frequencies, the used tags and other aspects of two snapshots taken from the Delicious system. This work was the first to explore user behavior in these systems and showed that users in Delicious exhibit a variety of different behaviors. Furthermore, the authors elaborated on different functions tags have for bookmarks. Examples are:

[18] http://www.connotea.org

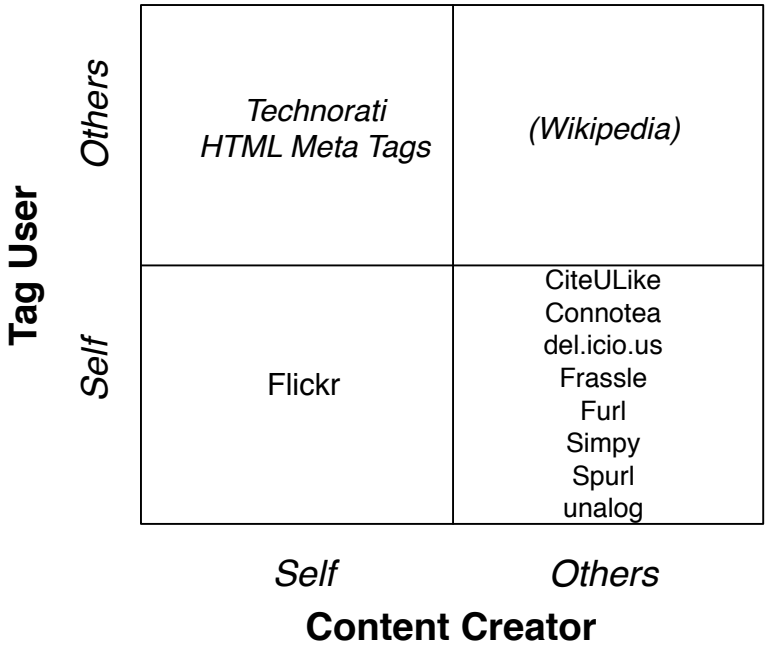


Figure 2.2.: Overview of motivators for tagging according to [Hammond et al., 2005]. The x-axis denotes the creators of content (own content vs content provided by others) whereas the y-axis shows for whom users annotate those resources (for themselves vs for others). Social bookmarking tools reviewed in the article are marked with plain text.

- *Identification what a resource is* - Specifies the kind of thing the annotated object is.
- *Identification what a resource is about* - Describing the topics of the document.
- *Indicators for qualities and characteristics* - Expressing the user's opinion of the resource.
- *Self Reference* - Tags beginning with "my" such as "`myown`" and "`mystuff`" that indicate a relation to the annotator.
- etc.

The authors further discover stability in relative proportions of tags within URLs. A possible explanation for this phenomenon is imitation and knowledge sharing processes occurring in tagging systems.

[Kipp and Campbell, 2006] investigate the Delicious system for inconsistencies and patterns in the included folksonomy to shed light on how tags (called *descriptors* by the authors) differ from descriptors applied by professional indexers. The authors identify so called *temporal tags* such as "`to_read`" and "`GTD`" (short for "getting things done") which cannot easily be integrated in currently available thesauri and give insight into the user perspective of such systems.

[Shen and Wu, 2005] analyze Delicious from a network theoretic perspective and report scale-free and small world properties within the fabric of this social system.

2.2.2. Studying tagging behavior

The identification and analysis of different types of tagging behavior is an ongoing subject in research.

[Xu et al., 2006] created a taxonomy of tags for the creation of a tag recommendation algorithm. These five categories are:

1. *Content-based* tags - give insight into the content or the categories of an annotated object (e.g. names, brands etc.)
2. *Context-based* tags - show the context under which the resource is stored (examples are: location, time, etc.)
3. *Attribute* tags - tell about properties of a resource
4. *Subjective* tags - explain the user's opinion of a given resource such as "`informative`" or "`cool`".
5. *Organizational* tags - enable a user to organize her library (example tags are "`to_read`", "`myown`" etc.)

In addition, the authors establish criteria for "good" tags. They argue that *specific* tags are well suited to differentiate a resource efficiently but

cannot be used for object discovery while broad or *generic* tags support the navigation and do not help to narrow down resources.

[Coates, 2005] hypothesizes two distinct tagging approaches. The first approach treats tags as a replacement for folders. This way tags describe the category where the annotated resource belongs to. The other approach simply uses tags on resources which make sense to the user and characterize the resource in a detailed manner.

[Farooq et al., 2007] propose six tag metrics such as tag growth, tag reuse, tag frequency, tag obviousness etc. to evaluate tagging behavior in folksonomic systems and evaluate them on a snapshot from CiteULike. Based on this analysis, the authors suggest three design heuristics for the CiteSeer scholarly digital library system:

- *User interfaces should facilitate the reuse of tags* - The analysis shows that users in CiteULike tend not to reuse tags by other users, but steadily introduce their own new tags into the system. For this reason the authors present three categories of tags to which a tagged publication can have been assigned to:
  - **Global Tags** - Tags which were used by all users of the social system.
  - **Personal Tags** - Tags that were used previously by the user.
  - **Paper-specific Tags** - Tags that all users of the system assigned to the respective paper.
- *Tags that have a high informational value should be used for recommendation* - By tags "that have a high informational value" the authors specify tags which are discriminative as well as non-obvious.
- *Support tagging season periods with corresponding scholarly resources* - The users should be supported in other activities related to tagging by supplementing other resources such as relevant publications.

The authors point out that these metrics might be useful for domains different from the scholarly area. Furthermore, they raise the question of how interpretable the different metrics are in other areas and argue

that different measures might be of different importance depending on the particular domain in question.

[Heckner et al., 2008] perform a comparative study on four different tagging systems (YouTube, Connotea, Delicious and Flickr) and observe differences in tagging behavior for different digital resources. General trends the authors identify were amongst others:

- Photos are tagged for content.
- Photos are tagged for location.
- Videos are tagged for persons.
- Scientific articles are tagged for time and task.

However, it is observed that caution has to be applied when interpreting these trends since different platforms entail varying objectives and motivations. While a user uploads a video to YouTube in order to make it accessible to others, systems like Connotea might serve a personal informational management purpose. The authors also examine the behavior of *overtagging* found in YouTube where content creators tag their resources excessively in order to make their resources better found by other users. This excessive tagging is done by copying a video's title or transcript into the tags. Also *tag avoidance* is a property observed in tagging systems such as Flickr since additional organizational structures (such as photo sets) exist and tagging is a supplementary feature.

Another work by [Heckner et al., 2009] illustrates a survey on 142 users of Flickr, YouTube, Flickr and Connotea. The study concludes a variety of different user intentions and tag usage scenarios with resource sharing being the top identified incentive. For this purpose the authors propose a segmentation of intentions for tagging into two different areas: *personal information management* (in short *PIM*) and *resources sharing*.

Figure 2.3 shows the model for information behaviors in social tagging systems which includes different types of users with their documents and the associated interaction processes.

[Rader and Wash, 2008] study the influences of users' tag choices in Delicious using Logistic Mixed Regression to answer the following ques-

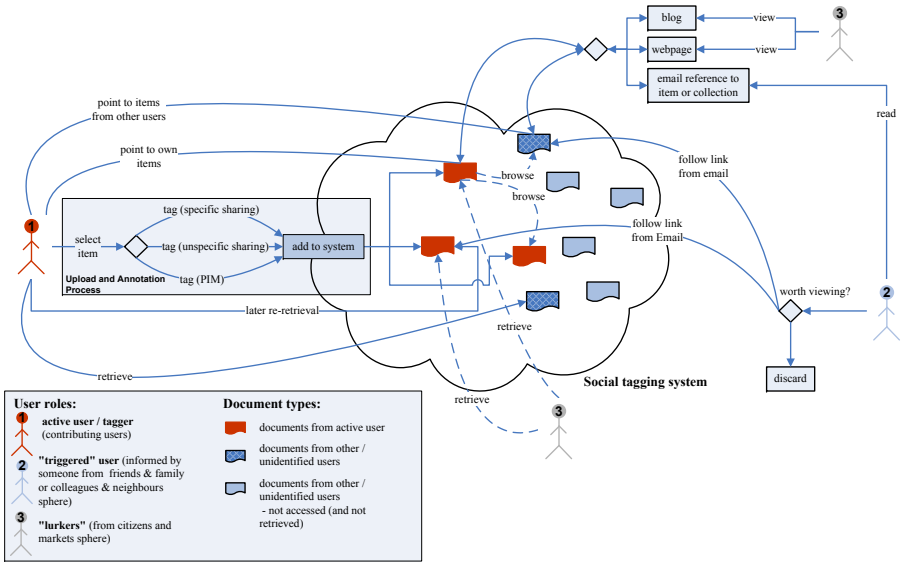


Figure 2.3.: Model of information behaviors as introduced by [Heckner et al., 2009]

tions:

- Do users imitate tags that have been applied previously by other users to the same resource?
- Do users re-use tags they have previously used on other resources?
- Do users annotate resources by using tags provided by the Delicious recommendation interface?

The work shows that users of the Delicious system are more focused on their personal information management than establishing a shared collective vocabulary. The authors argue that if this is the case users in these systems may find tags not to be useful for search and retrieval.

[Sen et al., 2006] introduce a user-centered model for vocabulary evolution to answer the questions of how strongly a user's investment and habit as well as the community of a system influence tagging behavior. The authors combine the seven tag classes by [Golder and Huberman, 2006] into three classes of their own:

- *Factual Tags* - Properties (like place, time, etc.) of the movie.
- *Subjective Tags* - Users' opinions about the annotated movie.
- *Personal Tags* - Tags which are used to organize a user's library (e.g. "`to_watch`")

To evaluate the classes human subjects labelled 3,263 tags into the three categories. An additional "other" class was used to signal if a tag was not understandable or did not fit into a single class. In total the annotators reached agreement on 87% of the tags which were 63% factual, 29% subjective, 3% personal and the rest (5%) other. Additional results were that investment and habit indeed influence tag applications and the influence increases the more users apply tags to resources. In another experiment it was found that the community has a large impact on the tagging process especially if a user has seen a lot of tags before starting the tagging endeavor. A survey of users of MovieLens[19] evaluated the usefulness of the three tagging classes on the following five tasks:

- Self-expression
- Organizing
- Learning
- Finding
- Decision support

The results of this survey conclude that different tag classes serve different tasks. To give examples: Subjective tags are useful for self-expression and factual tags can be used for learning additional information about movies.

A case study by [Wash and Rader, 2007] focuses on the incentives users have when they utilize a social computing systems. By analyzing interviews conducted with twelve users from the Delicious system the authors got the following top answers for the heuristics of choosing a tag:

- Reuse of tags that were applied to other resources before.

[19] http://movielens.umn.edu/

- Create and adhere to mental rules or definitions for specific tags.
- Choose terms she or he expects to search on.

The three top reasons for using the Delicious bookmarking facilities were the following:

- Keeping track of useful or interesting web sites.
- To access bookmarks from multiple computers.
- To achieve recognition from other users in the system.

In addition, the authors identify two different roles for users: Information producers and seekers who are browsing and searching within the system. The incentives also have to be aligned with these two roles.

[Marlow et al., 2006] show two high level types of categorization for motivation: *organizational* and *social* practices. The authors further elaborate a list of incentives by which users can be motivated:

- *Future retrieval* - Using tags to make them easier to find by the annotator himself.
- *Contribution and sharing* - Applying keywords in order to create clusters of resources to make them retrievable by other users.
- *Attract attention* - Driving other users towards own resources by using shared tagging vocabulary.
- *Play and competition* - Applying tags based on a game's rule set (such as the ESP game).
- *Self representation* - Putting a personal signature on resources.
- *Opinion expression* - Applying tags to express value judgement.

In a preliminary case study of Flickr the authors apply their presented usage model and show that the Flickr system is different from Delicious in terms of tag usage.

Work by [Nov et al., 2009] explores the range of different factors (motivational, structural and tenure) which can be observed on the social photo-sharing platform Flickr. The authors present a research model which

enables the application of motivation theories on data from an online social community. An individual's motivations which are investigated are *enjoyment*, *commitment*, *self development* and *reputation*. The included study uses a combination of signals from the system as well as a questionnaire in order to evaluate what drives users to contribute to these social platforms. An interesting aspect the authors find is that users committed to self development tend to share less resources per year than users driven by other incentives. This indicates that these users are more focused on higher quality and therefore more cautious in their photo sharing endeavor.

We can see that the presented work on analyzing tagging behavior either uses expert judgement or direct user questionnaire for the evaluation of user behavior and underlying tagging motivation. The work shown in this thesis introduces an automatic way to evaluate user behavior to get insight into a user's motivations.

2.2.3. Efficiency and navigability of social tagging systems

Another topic of research relevant for this dissertation is recent work about efficiency of tags and the navigability in tagging systems.

[Chi and Mytkowicz, 2008] investigate Delicious from an information theory perspective to evaluate the effectiveness of resource encoding via tags. They show measures to monitor this effectiveness and find that with increasing popularity the benefit of tags is dwindling due to "the rich get richer" phenomenon that occurs with popular tags.

As described in Section 1.2.2, tag clouds are often used to visualize vocabulary of resources or libraries in social tagging systems. [Helic et al., 2010] examine the usefulness of tag clouds as a tool for navigation from a network theoretic perspective. The included analysis shows that theoretically, tag-resource networks are effectively navigable. However, in reality they are negatively influenced by interface restrictions like for example pagination.

In order to investigate if social tagging systems are pragmatically useful for navigation [Helic et al., 2011] establish a framework that uses decen-

tralized search with hierarchies generated from the folksonomy as background knowledge. The authors find that the usefulness of four different state of the art tag hierarchy construction algorithms for navigation vary significantly.

In a follow-up paper [Helic and Strohmaier, 2011] study the utility of tag hierarchies induced from social tagging data for navigational aspects opposed to ontological usefulness in previous work. Their work shows that existing algorithms for the hierarchy construction perform poorly in the light of navigational tasks. Informed by these results the authors present an adaptation of a tag hierarchy algorithm by [Benz et al., 2010] that outperforms existing approaches.

2.2.4. Models for the analysis of social tagging systems

This section gives insight into work that has been done in modeling dynamics and behavior in folksonomic systems:

[Halpin et al., 2007] introduce a generative model to analyze the dynamics of collaborative tagging systems based on [Mika, 2007] that additionally incorporates aspects of information value for tags. Using this model the authors show that tagging distributions tend to stabilize to power law distributions over time. Further, they empirically examine the tagging history of web sites to analyze the dynamics found in collaborative tagging. Using the introduced methods it is possible to detect at what point a tagged resource has stabilized to power law.

[Cattuto et al., 2007] establish a stochastic model for user tagging behavior based on two aspects found in collaborative tagging systems: A frequency based approach based on the notion that users are exposed to each others' tagging activity and Yule-Simon's heavy-tailed memory model to mimic the idea of resources aging in the system. Even though the presented model has simple characteristics it is able to reproduce features found on a dataset that was analyzed in previous work.

[Dellschaft and Staab, 2008] present a stochastic dynamic model for simulating the evolution of tag streams in social tagging systems. This model distinguishes two possibilities for a user to assign tags: She either imitates

activity previously seen in the tag stream or chooses words from her own active vocabulary for tag assignments. The authors argue that previous models are only partially successful in the simulation of tag streams due to the omission of the user's background knowledge and imitate this aspect by incorporating a probability estimation stemming from a web corpus. An interesting observation in their evaluation finds that the imitation rate during tag assignments ranges between 60% and 90%.

2.2.5. Cognitive aspects of tagging

An important aspect when studying social tagging systems are the cognitive processes that occur during the course of tagging individual resources and using these systems. The following gives a short overview of work dealing with cognitive aspects of folksonomies:

[Sinha, 2005] conducts a cognitive analysis on tagging and relates it to the process of categorization. The author describes tagging as a one stage process consisting of the activation of concepts related to the object in question. The subsequent recording of these concepts can be seen as tagging. Categorization on the other hand requires a decision on one or more of these concepts increasing the cognitive cost of this process. Figure 2.4 and 2.5 show graphical illustrations provided by the author to visualize these differences.

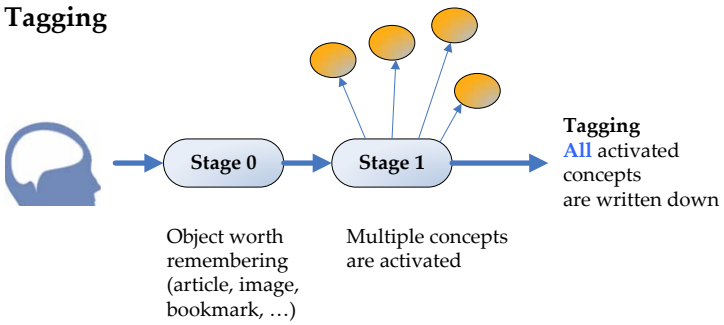


Figure 2.4.: Process of tagging according to [Sinha, 2005]

[Hong et al., 2008] introduce the social tagging application *SparTag.us*. The system employs a technique called *Click2Tag* to provide in-place, low

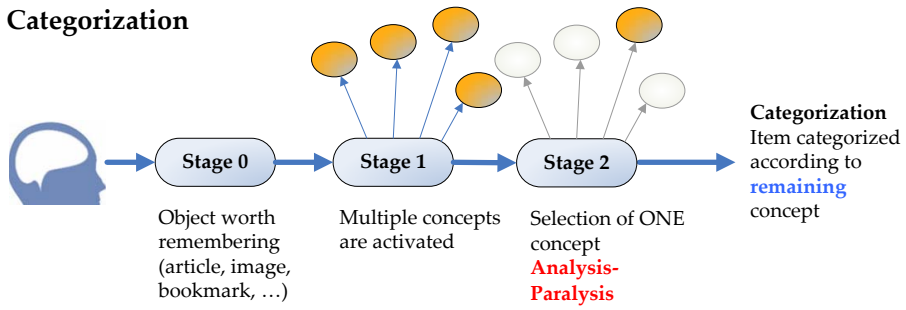


Figure 2.5.: Process of categorization according to [Sinha, 2005]

cost tagging for web documents. The system allows tagging on the fly by either clicking on words found in the text of the document or supplying additional tags by entering them in a text field. The authors find tagging by clicking to generate bottom-up, content-driven annotations whereas freely chosen tags lead to top-down, knowledge-driven annotations.

In another work [Ley and Seitlinger, 2010] argue that the analysis of emergent semantics in folksonomies also needs to address how users process information cognitively. The authors hypothesize that over time users in social tagging systems will undergo a shift in the *basic level advantage*[20] and subsequently use more specific categories/tags. The experimental study however found contradicting results: User groups who had less time to establish a shared understanding shifted to more specific levels than the other user groups.

[20]The basic level advantage describes the preferred abstraction level in a taxonomy that is used when a user has to classify objects of the real world.

3. Papers

This cumulative dissertation consists of five main publications detailing the path of research taken for the analysis of tagging motivation and a supplementary publication that shows a step towards future research. Figure 1.4 gives an illustration how the publications relate to the research questions and the topics that are covered in this thesis.

3.1. Main publications

The main papers of this dissertation that span the research on the aspects of user motivation in social tagging systems are the following:

1. Strohmaier, Markus; Körner, Christian and Kern, Roman (2010). *Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems.* In Proceedings of the International AAAI Conference on Weblogs and Social Media.

2. Kern, Roman; Körner, Christian and Strohmaier, Markus (2010). *Exploring the Influence of Tagging Motivation on Tagging Behavior.* In Proceedings of the 14$^{\text{th}}$ European Conference on Research and Advanced Technology for Digital Libraries, ECDL'10, pages 461–465, Berlin, Heidelberg. Springer-Verlag.

3. Körner, Christian; Kern, Roman; Grahsl, Hans-Peter and Strohmaier, Markus (2010b). *Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation.* In 21$^{\text{st}}$ ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada. ACM.

4. Körner, Christian; Benz, Dominik; Strohmaier, Markus; Hotho, An-

dreas and Stumme, Gerd (2010a). *Stop Thinking, Start Tagging - Tag Semantics Emerge From Collaborative Verbosity.* In Proceedings of the 19th International World Wide Web Conference (WWW 2010), Raleigh, NC, USA. ACM.

5. Zubiaga, Arkaitz; Körner, Christian and Strohmaier, Markus (2011). *Tags vs Shelves: From Social Tagging to Social Classification.* In Proceedings of the 22nd ACM conference on Hypertext and Hypermedia (HT 2011), pages 93–102, New York, NY, USA. ACM.

3.2. Additional publication

In addition to the main papers, this dissertation also includes a publication evaluating different measures for tag abstractness in social tagging systems. This work is a step towards a future distinction of users into two orthogonal categories of motivation - the notion of *generalists* and *specialists*.

- Benz, Dominik; Körner, Christian; Hotho, Andreas; Stumme, Gerd, and Strohmaier, Markus (2011). *One Tag to Bind Them all: Measuring Term Abstractness in Social Metadata.* In Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Pan, J., and Leenheer, P. D., editors, Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Crete.

3.3. Contributions to the papers

The following section details the contributions of other researchers and the author to the papers presented in this thesis.

- Strohmaier, Markus; **Körner, Christian** and Kern, Roman (2010). *Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems.* In Proceedings of the International AAAI Conference on Weblogs and Social Media.

- Kern, Roman; **Körner, Christian** and Strohmaier, Markus (2010). *Exploring the Influence of Tagging Motivation on Tagging Behavior.* In Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries, ECDL'10, pages 461–465, Berlin, Heidelberg. Springer-Verlag.

- **Körner, Christian**; Kern, Roman; Grahsl, Hans Peter and Strohmaier, Markus (2010b). *Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation.* In 21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada. ACM.

The original idea of categorizers and describers as well as the measures included in the papers originate from lengthy discussions between all authors. The design of the experiments and evaluation of the results stem from extensive discussions between all participating authors. The datasets were in part generated by the author and in part crawled by Hans Peter Grahsl during the course of his Master's thesis (cf. [Grahsl, 2010]). The design and setup of the included user study was developed by Hans-Peter Grahsl and the author. Further the tag agreement study was conducted by the author.

- **Körner, Christian**; Benz, Dominik; Strohmaier, Markus; Hotho, Andreas and Stumme, Gerd (2010a). *Stop Thinking, Start Tagging - Tag Semantics Emerge From Collaborative Verbosity.* In Proceedings of the 19th International World Wide Web Conference (WWW 2010), Raleigh, NC, USA. ACM.

The analysis of user behavior for this paper was done by the author of this thesis. The utilities to measure semantic relatedness were provided by members of KDE Kassel (Dominik Benz, Andreas Hotho, Gerd Stumme). The design of the experiments exploring the implications of user behavior on emergent semantics was developed in equal parts by them and the author.

- Zubiaga, Arkaitz; **Körner, Christian** and Strohmaier, Markus (2011). *Tags vs Shelves: From Social Tagging to Social Classification.* In Proceedings of the 22nd ACM conference on Hypertext

and hypermedia, HT '11, pages 93–102, New York, NY, USA. ACM.

The framework for the classification of documents into categories was provided by Arkaitz Zubiaga. The analysis of user behavior of GoodReads and LibraryThing was done by the author of this thesis. The story, design of the experiments and the elaboration of the results originate from extensive discussions between all participating authors during the stay of Arkaitz Zubiaga in Graz.

- Benz, Dominik; **Körner, Christian**; Hotho, Andreas; Stumme, Gerd and Strohmaier, Markus. (2011). *One Tag to Bind Them all: Measuring Term Abstractness in Social Metadata*. In Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Pan, J., and Leenheer, P. D., editors, Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), Heraklion, Crete.

The network-theoretic measures for abstractness and the evaluational setup were developed in equal parts by members of KDE Kassel (Dominik Benz, Andreas Hotho and Gerd Stumme) and the author of the thesis beforehand. The experimental design for this paper and the interpretation of the results originate from extensive virtual meetings and email correspondence between all authors.

The author of this thesis was involved in the measures' development, validation and experimentation throughout the series of research on tagging motivation under the supervision of Markus Strohmaier.

The following sections give an overview and state the contributions of each paper.

3.4. Why Do Users Tag? - Detecting Users' Motivation for Tagging in Social Tagging Systems

This paper is the entry point to the research on tagging motivation as it introduces a distinction between two types of tagging motivation - *Categorizers* and *Describers* - and shows differences between the two different kinds of user groups. Categorizers are users driven by the effort of cataloging their documents in a personal and consistent manner, who avoid synonyms and reuse tags. Their primary use of tags is the later browsing of their own personomy. Categorizers stand in contrast to describers who try to annotate their resources in a multifaceted and descriptive way. In comparison, they use a open vocabulary of objective tags and want to enable later retrieval by themselves and others.

For the differentiation of the two types of tagging motivation we present measures that evaluate personomies to detect description as well as categorization behavior. The key advantages of the presented measures are that they are *content agnostic*, *language independent* and based on *individual user characteristics*.

In an experiment we apply the measures to *synthetic* (a Flickr and ESP Game snapshot[21]) and *real-world* personomy datasets (Delicious, BibSonomy, CiteULike and MovieLens) and find that tagging motivation varies within and across a variety of different social tagging systems. In an additional experiment we analyze the influence tagging motivation of the individual user groups has on the underlying tag agreement and find that describers agree in general on more tags than their counterparts who are driven by categorization.

[21] http://www.gwap.com/gwap/

Why do Users Tag?
Detecting Users' Motivation for Tagging in Social Tagging Systems

Markus Strohmaier
Graz University of Technology
and Know-Center
Inffeldgasse 21a, A-8010 Graz
markus.strohmaier@tugraz.at

Christian Körner
Graz University of Technology
Inffeldgasse 21a, A-8010 Graz
christian.koerner@tugraz.at

Roman Kern
Know-Center
Inffeldgasse 21a, A-8010 Graz
rkern@know-center.at

Abstract

While recent progress has been achieved in understanding the structure and dynamics of social tagging systems, we know little about the underlying user motivations for tagging, and how they influence resulting folksonomies and tags. This paper addresses three issues related to this question: 1.) What motivates users to tag resources, and in what ways is user motivation amenable to quantitative analysis? 2.) Does users' motivation for tagging vary within and across social tagging systems, and if so how? and 3.) How does variability in user motivation influence resulting tags and folksonomies? In this paper, we present measures to detect whether a tagger is primarily motivated by categorizing or describing resources, and apply the measures to datasets from 8 different tagging systems. Our results show that a) users' motivation for tagging varies not only across, but also within tagging systems, and that b) tag agreement among users who are motivated by *categorizing resources* is significantly lower than among users who are motivated by *describing resources*. Our findings are relevant for (i) the development of tag recommenders, (ii) the analysis of tag semantics and (iii) the design of search algorithms for social tagging systems.

Introduction

A question that has recently attracted the interest of our community is whether the properties of tags in tagging systems and their usefulness for different purposes can be assumed to be *a function of the taggers' motivation or intention behind tagging* (Heckner, Heilemann, and Wolff 2009). If this was the case, the motivation for tagging (why users tag) would have broad implications. In order to assess the general usefulness of algorithms that aim to - for example - capture knowledge from folksonomies, we would need to know whether these algorithms produce similar results across user populations driven by different motivations for tagging. Recent research already suggests that different tagging systems afford different motivations for tagging (Heckner, Heilemann, and Wolff 2009), (Hammond et al. 2005). Further work presents anecdotal evidence that even within the same tagging system, the motivation for tagging between individual users may vary greatly (Wash and Rader 2007). Given these observations, it is interesting to study whether and how the analysis of user motivation for tagging is amenable to quantitative investigations, and whether folksonomies and their tags are influenced by different tagging motivations.

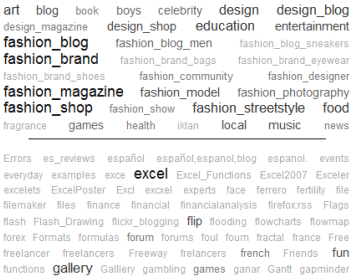


Figure 1: Examples of tag clouds produced by users who are driven by different motivations for tagging: categorization (top) vs. description (bottom)

Categorizing vs. Describing Resources

Tagging motivation has remained largely elusive until the first studies on this subject have been conducted in 2006. At this time, the work by (Golder and Huberman 2006) and (Marlow et al. 2006) have made advances towards expanding our theoretical understanding of tagging motivation by identifying and classifying user motivation in tagging systems. Their work was followed by studies proposing generalizations, refinements and extensions to previous classifications (Heckner, Heilemann, and Wolff 2009). An influential observation was made by (Coates 2005) and elaborated on and interpreted in (Marlow et al. 2006), (Heckner, Heilemann, and Wolff 2009) and (Körner 2009). This line of work suggests that a distinction between at least two types of user motivation for tagging is important: On one hand, users who are motivated by categorization view tagging as a means to *categorize resources* according to some high-level characteristics. These users tag because they want to construct and maintain a navigational aid to the resources for later browsing. On the other hand, users who are motivated by description view tagging as a means to accurately and

	Categorizer (C)	Describer (D)
Goal	later browsing	later retrieval
Change of vocabulary	costly	cheap
Size of vocabulary	limited	open
Tags	subjective	objective

Table 1: Differences between categorizers and describers

precisely *describe resources*. These users tag because they want to produce annotations that are useful for later searching. Figure 1 illustrates this distinction with tag clouds of actual users.

A distinction between categorizers and describers has been found to be important because, for example, tags assigned by describers might be more useful for information retrieval and knowledge acquisition (because these tags focus on the content of resources) as opposed to tags assigned by categorizers, which might be more useful to capture a rich variety of possible interpretations of a resource (because they focus on user-specific views on resources).

Table 1 illustrates a number of intuitions about the two identified types of tagging motivation. While these two categories make an ideal distinction, tagging in the real world is likely to be motivated by a combination of both. A user might maintain a few categories while pursuing a description approach for the majority of resources and vice versa, or additional categories might be introduced over time. In addition, the distinction between categorizers and describers is not about the semantics of tagging, it is a distinction based on the motivation for tagging. One implication of that is that it would be plausible for the same tag (for example '`java`') to be used by both describers and categorizers, and serve both functions at the same time. In other words, the same tag might be used as a category or a descriptive label.

In this paper, we are adopting the distinction between categorizers and describers to study the following research questions: 1) How can we measure the motivation behind tagging? 2) How does users' motivation for tagging vary across and within different tagging systems? and 3) How does tagging motivation influence resulting folksonomies?

Datasets And Experimental Setup

To study these questions, we develop a number of measures and apply them to a large set of personomies (i.e. complete tagging records of individual users) that exhibit different tagging behavior. We apply all measures to various tagging datasets. Then, we analyze the ability of measures to capture predicted (synthetic) behavior. Finally, we relate our findings to results reported by previous work.

Assuming that the different motivations for tagging produce different personomies (different tagging behavior over time), we can use synthetic data from extreme categorizers and describers to find upper and lower bounds for the behavior that can be expected in real-world tagging systems.

Dataset	$\|U\|$	$\|T\|$	$\|R\|$	$\|R_u\|_{min}$	$\|T\|/\|R\|$
ESP Game*	290	29,834	99,942	1,000	0.2985
Flickr Sets*	1,419	49,298	1,966,269	500	0.0250
Delicious	896	184,746	1,089,653	1,000	0.1695
Flickr Tags	456	216,936	965,419	1,000	0.2247
Bibsonomy Bookmarks	84	29,176	93,309	500	0.3127
Bibsonomy Publications	26	11006	23696	500	0.4645
CiteULike	581	148,396	545,535	500	0.2720
Diigo Tags	135	68,428	161,475	500	0.4238
Movielens	99	9,983	7,078	500	1.4104

Table 2: Overview and statistics of social tagging datasets. The asterisks indicate synthetic personomies of extreme categorization/description behavior.

Synthetic Personomy Datasets

To simulate behavior of users who are mainly driven by description, data from the ESP game dataset[1] was used. This dataset contains a large number of inter-subjectively validated, descriptive tags for pictures useful to capture describer behavior. To contrast this data with behavior of users who are mainly driven by categorization, we crawled data from Flickr, but instead of using the tags we used information from users' *photo sets*. We consider each photo set to represent a tag assigned by a categorizer for all the photos that are contained within this set. The personomy then consists of all photos and the corresponding photo sets they are assigned to. We use these two synthetic datasets to simulate behavior of "artificial" taggers who are mainly motivated by description and categorization.

Real-World Personomy Datasets

In addition to the synthetic datasets, we also crawled data from popular tagging systems. The datasets needed to be *sufficiently large* in order to enable us to observe tagging motivation across a large number of users and they needed to be *complete* because we wanted to study a users complete tagging history over time - from the users first bookmark up to the most recent bookmarks. Because many of the tagging datasets available for research focus on sampling data on an aggregate level rather than capturing complete personomies, we had to acquire our own datasets. An overview of the datasets is given in Table 2.

Detecting Tagging Motivation

While we have experimented with a number of measures, in the following we will present two measures that are capable of providing useful insights into the fabric of tagging motivation in social tagging systems. The measures introduced below focus on statistical aspects of users' *personomies only* instead of analyzing entire folksonomies.

Detecting Categorizers: The activity of tagging can also be viewed as an encoding process, where tags encode information about resources. If this would be the case, users

[1] http://www.cs.cmu.edu/~biglou/resources/

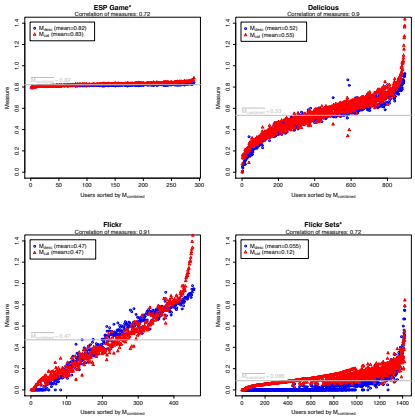


Figure 2: M_{desc} and M_{cat} at $|R_u| = 500$ for 4 datasets, including Pearson correlation and the mean value for $M_{combined}$.

motivated by categorization could be characterized by their encoding quality, where categorizers would aim to maintain high information value in their tag vectors. This intuition can be captured with the conditional entropy of $H(R|T)$, which will be low if the tag distribution efficiently encodes the resources. For normalization purposes, we relate the conditional entropy to an optimal encoding strategy given by the number of tags, resources and average number of tags per resource: $M_{cat} = \frac{H(R|T) - H_{opt}(R|T)}{H_{opt}(R|T)}$.

Detecting Describers: Users who are primarily motivated by description would generate tags that closely resemble the content of the resources. As the tagging vocabulary of describers is not bounded by taxonomic constraints, one would expect describers to produce a high number of unique tags - $|T|$ - in relation to the number of resources - $|R|$. One way to formalize this intuition is the *orphan ratio*, a measure capturing the extent to which a user exhibits description behavior: $M_{desc} = \frac{|\{t:|R(t)| \leq n\}|}{|T|}, n = \lceil \frac{|R(t_{max})|}{100} \rceil$

Both measures aim to capture different intuitions about using tags for categorization and description purposes. A combination of these measures - $M_{combined}$ - can be defined as their arithmetic mean: $M_{combined} = \frac{M_{desc} + M_{cat}}{2}$

Results and Discussion

The introduced measures have a number of useful properties: They are content-agnostic and language-independent, and they operate on the level of individual users. An advantage of content-agnostic measures is that they are applicable across different media (e.g. photos vs. text). Because the introduced measures are language-independent, they are applicable across different user populations (e.g. English vs. German). Because the measures operate on an personomy

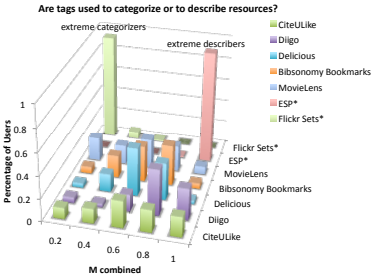


Figure 3: $M_{combined}$ at $|R_u| = 500$ for 7 different datasets, binned in the interval [0.0 .. 1.0]. The two back rows reflect opposite extreme behaviors.

level, only tagging records of individual users are required (as opposed to entire folksonomies).

Figure 2 depicts M_{desc} and M_{cat} measures for four tagging datasets at $|R_u| = 500$, i.e. at the point where all users have bookmarked exactly 500 resources. We can see that both measures identify synthetic describer behavior (ESP game, left top) and synthetic categorizer behavior (Flickr photosets, right bottom) as extreme behavior. We can use the synthetic data as points of reference for the analysis of real tagging data, which would be expected to lie in between these points of reference. The diagrams for the real-world datasets show that tagging motivation in fact mostly lies in between the identified extremes. The fact that the synthetic datasets act as approximate upper and lower bounds for real-world datasets is a first indication for the usefulness of the presented measures. We also calculated Pearson's correlation between M_{desc} and M_{cat}. The results, presented in Figure 2, are encouraging especially because the measures were independently developed based on different intuitions.

Figure 3 presents a different visualization for five selected tagging datasets. Each row shows the distribution of users for a particular dataset according to $M_{combined}$. Again, we see that the profiles of Delicious, Diigo and Movielens (depicted) as well as the other datasets (not depicted) largely lie in between these bounds. The characteristic distribution of different datasets provides first *empirical* insights into the fabric of tagging motivation in different systems, illustrating a broad variety within these systems.

In addition, we evaluated whether individual users that were identified as extreme categorizers / extreme describers by $M_{combined}$ were also confirmed as such by human subjects. In our evaluation, we asked one human subject (who was not related to this research) to classify 40 exemplary tag clouds into two equally-sized piles: a categorizer and a describer pile. The 40 tag clouds were obtained from users in the Delicious dataset, where we selected the top20 categorizers and the top20 describers as identified by $M_{combined}$. The inter-rater agreement kappa between the results of the human subject evaluation and $M_{combined}$ was 1.0. This means the human subject agrees that the top20 describers and the

k	10	20	30	40	50	60	70	80
Desc. Wins	379	464	471	452	380	287	173	69
Cat. Wins	56	11	5	7	5	3	4	4
Ties	65	25	24	41	115	210	323	427

Table 3: Tag agreement among Delicious describers and categorizers for 500 most popular resources. For all different k, describers produce more agreed tags than categorizers.

top20 categorizers (as identified by $M_{combined}$) are good examples of extreme categorization / description behavior. The tag clouds illustrated earlier (cf. Figure 1) were actual examples of tag clouds used in this evaluation.

In an additional experiment, we examined whether the intuition that describers agree on more tags is correct. For this purpose we divided the users of our Del.icio.us data set in groups of equal size. Users who had a $M_{combined}$ value lower than 0.5514 were referred to as *Delicious Categorizers* whereas users with a higher value were denoted *Delicious Describers*. For each of the two groups we generated a tag set of the 500 most popular resources. For both of these tag sets we calculated the tag agreement, i.e. the number of tags that k percent of users agree on for a given resource.

Table 3 shows the agreement values of k percent of users. We restricted our analysis to $T_u > 3$ in order to avoid irrelevant high values in this calculation. In all cases - for different values of k - describers produce more agreed tags than categorizers.

Conclusions

This paper introduced a quantitative way for measuring and detecting the tacit nature of tagging motivation in social tagging systems. We have evaluated these measures with synthetic datasets of extreme behavior as points of reference, via a human subject study and via triangulation with previous findings. Based on a large sample of users, our results show that 1) tagging motivation of individuals varies within and across tagging systems, and 2) that users' motivation for tagging has an influence on resulting tags and folksonomies. By analyzing the tag sets produced by Delicious describers and Delicious categorizers, we showed that agreement on tags among categorizers is significantly lower compared to agreement among describers. We believe that these findings have some interesting implications:

Usefulness of Tags: Our research shows that users motivated by categorization produce fewer descriptive tags, and that the tags they produce exhibit a lower agreement among users for given resources. This provides further evidence that not all tags are equally useful for different tasks, such as information retrieval. Rather the opposite seems to be the case: Without knowledge of users' motivation for tagging, an assessment of the usefulness of tags on a content-independent level seems challenging. The measures introduced in this paper aim to illuminate a path towards understanding user motivation for tagging in a quantitative, content-agnostic and language-independent way that is based on local data of individual users only. In subsequent work, the distinction between categorizers and describers was successfully used to demonstrate that emergent semantics in folksonomies are influenced by the users' population motivation for tagging (Körner et al. 2010).

Usage of Tagging Systems: While tags have been traditionally viewed as a way of freely describing resources, our analysis suggest that the motivation for tagging across different real world social tagging systems such as Delicious, Bibsonomy and Flickr varies tremendously. Moreover, our data shows that even within the same tagging systems the motivation for tagging varies strongly. The findings presented in this paper highlight several opportunities for designers of social tagging systems to influence user behavior. While categorizers could benefit from tag recommenders that recommend tags based on their individual tag vocabulary, describers could benefit from tags that best capture the content of the resources. Offering users tag clouds to aid the navigation of their resources might represent a way to increase the proportion of categorizers, while offering more sophisticated search interfaces and algorithms might encourage users to focus on describing resources.

Acknowledgments

Thanks to Hans-Peter Grahsl for support in crawling the data sets and to Mark Kroell for comments on earlier versions of this paper. This work is in part funded by the FWF Austrian Science Fund Grant P20269 TransAgere and the Know-Center Graz.

References

Coates, T. 2005. Two cultures of fauxonomies collide. *http://www. plasticbag. org/archives/2005/06/two_cultures_of_fauxonomies_collide/. Last access: May* 8:2008.

Golder, S., and Huberman, B. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32(2):198.

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

Heckner, M.; Heilemann, M.; and Wolff, C. 2009. Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In *ICWSM '09: Int'l AAAI Conference on Weblogs and Social Media*.

Körner, C.; Benz, D.; Strohmaier, M.; Hotho, A.; and Stumme, G. 2010. Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*. Raleigh, NC, USA: ACM.

Körner, C. 2009. Understanding the motivation behind tagging. ACM Student Research Competition - HT2009.

Marlow, C.; Naaman, M.; Boyd, D.; and Davis, M. 2006. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the 17th Conference on Hypertext and Hypermedia*, 31–40. New York, NY, USA: ACM.

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

3.5. Exploring the Influence of Tagging Motivation on Tagging Behavior

This poster paper gives additional information on the variety of categorizing and describing behavior in different social systems to deepen quantitative investigations into the reasons why users tag.

To evaluate the synthetic datasets for their usefulness as reference data, we compute the accuracy for the Flickr and the ESP Game snapshots respectively and achieved high values. Furthermore, we conducted a recommender evaluation to gain insight whether tagging motivation distinction has impact during the course of tagging and to find out which of the five measures captures the differentiation best. The evaluation concluded that describers tend to use tags which are similar to tags from other describers whereas categorizers stick to their own personal tag vocabulary.

Exploring the Influence of Tagging Motivation on Tagging Behavior

Roman Kern[1], Christian Körner[2], and Markus Strohmaier[1,2]

[1] Know-Center, Graz
[2] Graz University of Technology
rkern@know-center.at,{christian.koerner,markus.strohmaier}@tugraz.at

Abstract. The reasons why users tag have remained mostly elusive to quantitative investigations. In this paper, we distinguish between two types of motivation for tagging: While *categorizers* use tags mainly for categorizing resources for later browsing, *describers* use tags mainly for describing resources for later retrieval. To characterize users with regard to these different motivations, we introduce statistical measures and apply them to 7 different real-world tagging datasets. We show that while most taggers use tags for both categorizing and describing resources, different tagging systems lend themselves to different motivations for tagging. Additionally we show that the distinction between describers and categorizers can improve the performance of tag recommendation.

1 Introduction

Tags in social tagging systems are used for a variety of purposes [1]. In this paper, we study the distinction between two different tagging behaviors. The first type of tagging is similar to assign resources to a predefined classification scheme. Users motivated by this behavior use tags out of a controlled and closed vocabulary. These users, named *categorizers*, tag because they want to construct and maintain a navigational aid to resources for later browsing. On the other hand, users who are motivated by description view tagging as a means to accurately and precisely describe resources. Tags produced by this user group resemble keywords that are useful for later searching [2]. This distinction can be exploited for example to improve the performance of tag recommender systems and information retrieval applications. Figure 1 contrasts a tag cloud of a typical categorizer with a tag cloud of a typical describer.

2 Development of Measures

To characterize the extent to which users categorize or describe resources, we present statistical and information-theoretic measures that are independent of the meaning of tags, the language of tags, or the resources being tagged.

Characterizing Categorizers: The activity of tagging can also be viewed as an encoding process, where tags encode information about resources. If this

Fig. 1. Examples of tag clouds of a typical *categorizer* (left) and a typical *describer* (right)

would be the case, users motivated by categorization could be characterized by their encoding quality, where categorizers would aim to maintain high information value in their tag vectors. This intuition can be captured with the conditional entropy of $H(R|T)$, which will be low if the tag distribution efficiently encodes the resources, with R being the set of resources and T the set of tags. For normalization purposes, we relate the conditional entropy to an optimal encoding strategy given by the number of tags, resources and average number of tags per resource: $M_{cat} = \frac{H(R|T) - H_{opt}(R|T)}{H_{opt}(R|T)}$.

Characterizing Describers: Users who are primarily motivated by description would generate tags that closely resemble the content of the resources. As the tagging vocabulary of describers is not bounded by taxonomic constraints, one would expect describers to produce a high number of unique or very rare tags in relation to the number of resources. One way to formalize this intuition is the orphan ratio, a measure capturing the extent to which a user exhibits description behavior: $M_{desc} = \frac{|\{t:|R(t)| \leq n\}|}{|T|}, n = \lceil \frac{|R(t_{max})|}{100} \rceil$

A combination of these measures - $M_{combined}$ - can be defined as the arithmetic mean of M_{desc} and M_{cat}. A detailed description of the measures and a comparison with other measures can be found in [3].

3 Application of Measures

Synthetic Datasets: As a first check of they usefulness of the measures we first applied them on two synthetic datasets, which are designed to resemble extreme categorizing and describing behavior. The synthetic dataset for describer behavior is based on the ESP game dataset, where users describe images (290 users and 29,834 tags). For the extreme categorizers we used the photoset feature from Flickr, where users sort their pictures into albums, just like users would organize their pictures in folders on their hard drives (1,419 users with 39,298 tags). The accuracy with which a measure can identify ESP game data as describers, and Flickr Sets data as categorizers can act as an approximation of its validity. In this simplified setting, $M_{combined}$, M_{desc} and M_{cat} achieve high accuracy values of 99.94%, 99.82% and 99.94% respectively.

Real-World Datasets: We have gathered 7 real-world datasets from different social tagging systems. For each tagging system, we acquired data from users with a minimum number of resources $|R_u|_{min}$. See table 1 for an overview of the size of the datasets. We applied the $M_{combined}$ measure on all datasets to study whether the various tagging systems differ in regard to the two user types. Figure 2 demonstrates that the distribution of describers and categorizers vary

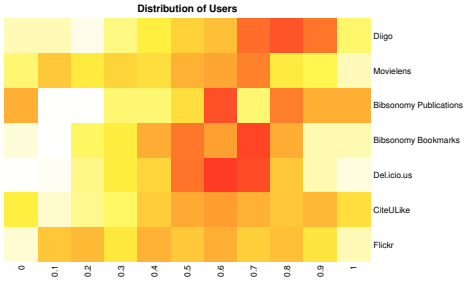


Fig. 2. Distribution of users of the real-world datasets according to the $M_{combined}$ measure. The intensity of the color encodes the relative number of users within a bin. The bins on the left side represent categorizer, while the rightmost bins represent users that display a behavior typical for describers. For example the `Flickr` dataset contains users evenly distributed between the two extremes, whereas the majority of users in the `Diigo` dataset are identified as describers.

between the individual social tagging systems. For example the `Diigo` dataset contains many users that are identified as describers. One possible reason for this might be the fact, that the `Diigo` platform not only offers the possibility to tag resources, but also to create so called bookmark lists, which is better suited to categorize resources.

Tag Recommender Finally we implemented two simple tag recommender systems to test whether the distinction between the two users groups could improve the performance of tag recommendation. The first recommender draws tags from the personal tagging history of a user and is labeled as *personomy-based recommender*. The *folksonomy-based recommender* suggests the most frequent tags as used by other describer users. Users were split into a describer and categorizer group according to the $M_{combined}$ measure. The baseline was produced by randomly assigning users to one of the two groups. For the evaluatcion we used the `Del.icio.us` dataset, as the folksonomy-based recommender requires resources tagged by multiple users. Figure 3 depicts the performance of the tag recommenders for different splits of the userbase (from 10% categorizers and 90% describers up to a 90%:10% split). One can see that using the personal tagging history is helpful for categorizers, while describers appear to tags similar to other users (describers) in the folksonomy. Especially of interest is the point where the relative improvement of the two recommenders intersect each other (right chart in figure 3). When developing a production tag recommender, this would be the point to switch from personomy-based tag recommendation for categorizers to a folksonomy-based recommender for describers.

Table 1. Overview of the size and characteristics of the crawled real-world datasets.

Dataset	$\|U\|$	$\|T\|$	$\|R\|$	$\|R_u\|_{min}$	$\|T\|/\|R\|$
Delicious	896	184,746	1,089,653	1,000	0.1695
Flickr Tags	456	216,936	965,419	1,000	0.2247
Bibsonomy Bookmarks	84	29,176	93,309	500	0.3127
Bibsonomy Publications	26	11006	23696	500	0.4645
CiteULike	581	148,396	545,535	500	0.2720
Diigo Tags	135	68,428	161,475	500	0.4238
Movielens	99	9,983	7,078	500	1.4104

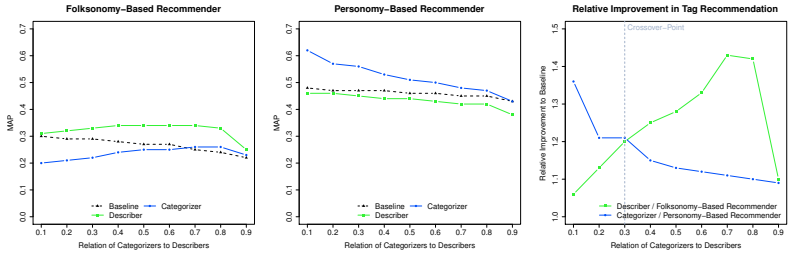


Fig. 3. Suggesting tags also used by other users appears to be a good strategy for describers (left). Categorizer prefer to reuse tags from their personal tagging history (middle). The relative improvements indicates that for about 30% of all users in our `Del.icio.us` dataset the personomy-based recommender is the better choice (right).

4 Conclusion

We showed that different tagging systems lend themselves to different motivations for tagging. Our results reveal that even within tagging systems, tags are adopted in different ways. One of the major implications of our work is that tagging motivation exhibits significant variety, which could play an important part in a range of problems including tag recommendation and information retrieval. In previous work [4], we have demonstrated that the motivation behind tagging influences the performance of semantic acquisition algorithms in folksonomies. Improving existing state-of-the-art tag recommenders by incorporating the tagging motivation is one of the main goals of our future work.

Acknowledgements: The research presented in this work is in part funded by the Know-Center and the FWF Austrian Science Fund Grant P20269. The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of VIT, Austrian Ministry of WA and by the State of Styria.

References

1. Heckner, M., Heilemann, M., Wolff, C.: Personal information management vs. resource sharing: Towards a model of information behaviour in social tagging systems. In: Int'l AAAI Conference on Weblogs and Social Media (ICWSM), San Jose, CA, USA (2009)
2. Strohmaier, M., Körner, C., Kern, R.: 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 23-26. (2010)
3. Körner, C., Kern, R., Grahsl, H.P., Strohmaier, M.: Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In: 21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010), Toronto, Canada, ACM (June 2010)
4. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Stumme, G.: Stop thinking, start tagging: Tag semantics arise from collaborative verbosity. Proceedings of the 19th Int'l Conf. on World Wide Web - WWW '10 (2010)

3.6. Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation

This paper defines and evaluates the previously introduced measures used to differentiate the two types of tagging motivation - categorizers and describers: The *Tag Resource Ratio* measures the ratio between tags and resources found in a user's library. The higher this ratio is the more likely a user is a describer expressing her annotations in a verbose way. The *Orphaned Tag Ratio* analyzes the proportion of a user's infrequently used tags. A categorizer establishes a set of tags which are reused and would therefore be reflected in a low score. The *Conditional Tag Entropy* examines how good a user applies tags to encode resources. A describer would not care to use the tag vocabulary as balanced as possible which would be seen in a high tag entropy. The fourth measure introduced is the *Overlap Factor* that quantifies how strongly resources are overlapping in tag sets created by the user. Categorizers might try to construct tag sets that are not overlapping in terms of documents found in the library. The last measure - the *Tag/Title Intersection Ratio* - indicates how likely a user assigns tags to a resource which are already contained in the title. Since a categorizer would stick to her own convention for the creation of the tagging vocabulary the probability that she assigns tags from the title would be low, leading to a low score for this ratio.

We perform an quantitative as well as a qualitative evaluation to characterize the different measures. For the qualitative evaluation we conducted an empirical user study to investigate the usefulness of the different measures for tagging motivation. Furthermore, we assessed the performance of the five measures by using accuracy as a metric and found the best performing measures to be Tag/Resource Ratio, Overlap Factor and Tag/Title Intersection Ratio.

Of Categorizers and Describers: An Evaluation of Quantitative Measures for Tagging Motivation

Christian Körner
Knowledge Management Institute
Inffeldgasse 21A
8010 Graz, Austria
christian.koerner@tugraz.at

Roman Kern
Know-Center
Inffeldgasse 21A
8010 Graz, Austria
rkern@know-center.at

Hans-Peter Grahsl
Graz University of Technology
Inffeldgasse 21A
8010 Graz, Austria
grahsl@student.tugraz.at

Markus Strohmaier
Knowledge Management Institute and Know-Center
Inffeldgasse 21A
8010 Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

While recent research has advanced our understanding about the structure and dynamics of social tagging systems, we know little about (i) the underlying motivations for tagging (why users tag), and (ii) how they influence the properties of resulting tags and folksonomies. In this paper, we focus on problem (i) based on a distinction between two types of user motivations that we have identified in earlier work: Categorizers vs. Describers. To that end, we systematically define and evaluate a number of measures designed to discriminate between describers, i.e. users who use tags for *describing resources* as opposed to categorizers, i.e. users who use tags for *categorizing resources*. Subsequently, we present empirical findings from qualitative and quantitative evaluations of the measures on real world tagging behavior. In addition, we conducted a recommender evaluation in which we study the effectiveness of each of the presented measures and found the measure based on the tag content to be the most accurate in predicting the user behavior closely followed by a content independent measure. The overall contribution of this paper is the presentation of empirical evidence that tagging motivation can be approximated with simple statistical measures. Our research is relevant for (a) designers of tagging systems aiming to better understand the motivations of their users and (b) researchers interested in studying the effects of users' tagging motivation on the properties of resulting tags and emergent structures in social tagging systems.

HT'10, June 13–16, 2010, Toronto, Ontario, Canada..

Categories and Subject Descriptors

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

General Terms

Algorithms, Human Factors, Measurement

Keywords

tagging, user motivation, measures, social software

1. INTRODUCTION

Social tagging systems, such as Flickr, Del.icio.us and others, have emerged as an interesting alternative for users to annotate and organize resources on the web. While past research has made significant advances towards understanding the complex dynamics and structure of tagging systems as a whole (cf. [4, 2, 9]), we know surprisingly little about the motivations of individual users, and why they tag. The motivation for tagging can be regarded as an important issue since recent research suggests that it has a direct influence on the properties of resulting tags and folksonomies [7, 10, 11]. If the intuitions were known to designers of social tagging platforms a number of current research questions would be easier to answer. Examples include enhancement in ontology learning as well as improving tag recommendation engines and the finding of suitable terms for search in these systems. However, the reasons why users tag - and ways to measure it - have remained largely elusive.

This paper aims to tackle the problem of tagging motivation identification by systematically deriving and evaluating a set of measures as an instrument for characterizing user motivation in social tagging systems. Valid measures for tagging motivation could act as a stepping stone for studying the different ways in which user motivation influences the properties of tags and the dynamic structures emerging in social tagging systems [7].

A number of different categories for tagging motivation have been proposed in the literature. In this paper we use a

simplified distinction identified by us in earlier work: *categorization* vs. *description* (cf. by [10], [19] and [11]). Users who are motivated by categorization view tagging as a means to *categorize resources* according to some shared high-level characteristics. Categorizers tag because they want to construct and maintain a navigational aid to the resources for later browsing. On the other hand, users who are motivated by description view tagging as a means to accurately and precisely *describe resources*. Describers tag because they want to produce annotations that are useful for later retrieval. This distinction has been found to be important because, for example, tags assigned by describers might be more useful for information retrieval (because these tags focus on the content of resources) as opposed to tags assigned by categorizers, which might be more useful to capture a rich variety of possible interpretations of a resource (because they focus on user-specific views on resources).

In this paper, we want to examine the usefulness of different measures for discriminating between *categorizers and describers*, a problem that we have started to formulate in previous research [10]. To that end, we will express different intuitions about this distinction and systematically derive a number of measures based on them. The presented paper makes the following contributions: (i) we introduce a number of measures for tagging motivation and corresponding intuitions (ii) we evaluate the introduced measures both qualitatively (in human subject studies) and quantitatively (in experiments) and (iii) provide results suggesting what measures are indicative of what kind of tagging motivation. The overall contribution of our paper is an increased understanding about measures aimed at capturing different aspects of tagging motivation. Our results are relevant for researchers interested in user motivation, adaptation and user behavior in social tagging systems.

The paper is organized as follows: In Section 2 we give an overview of related work. This is followed by section 3 which discusses the two types of tagging motivation and their characteristics. In section 4, a number of potential measures to distinguish categorizers from describers are introduced. The dataset and correlations between the measures are presented in Section 5. Qualitative and quantitative evaluations of the proposed measures are described in sections 6 and 7. Finally, in section 8 we summarize our findings and discuss conclusions for future work.

2. RELATED WORK

Relevant research on the motivation behind tagging is presented chronologically in Table 1. An interesting observation is that research on tagging motivation is shifting from anecdotal evidence (cf. [3, 5]) and theoretical grounding (cf. [4]) to larger datasets and empirical validation (cf. [20, 1, 6, 16]). We can also observe a lack of consensus about different categories of tagging motivation, evaluation strategies, and the anticipated scope that such studies should cover. While early work focused on conceptualizing tagging motivation, recent work lays more focus on quantitative aspects.

In our own work [10], we presented an initial attempt towards quantitative measures for tagging motivation, discriminating between categorizers and describers. Our preliminary results showed that tagging motivation not only varies between tagging systems, but that different users within the same tagging system also exhibit vast differences in the motivations for tagging. In [19], we elaborate measures to distinguish the two types of tagging motivation further and show that a particular property of tags in social tagging systems - tag agreement - is influenced by tagging motivation. In our most recent work [11], we found a link between the pragmatics of tagging (why and how users tag) and the resulting folksonomical structure.

In the paper at hand, we expand this line of research by systematically defining and evaluating a range of different measures for characterizing tagging motivation in social tagging systems. While effects of tagging motivation have been studied in different contexts, the measures were largely based on intuitions and validation of these measures has not received sufficient attention yet. As a consequence, we aim to address this gap by evaluating potential measures for tagging motivation both qualitatively (in human subject studies) and quantitatively (in experiments).

3. TYPES OF TAGGING MOTIVATION

Based on previous work [10, 19, 11], we differentiate between two particular types of tagging motivation – *categorizers* and *describers* – which can be characterized in the following way:

3.1 Using Tags to Categorize Resources

Users who are motivated by categorization use tags to construct and maintain a navigational aid to the resources they annotate. For this purpose, categorizers aim to establish a stable vocabulary based on their personal preferences and behavior. To keep navigation in this vocabulary as simple and non-redundant as possible, categorizers tend to avoid tags which have similar semantic meaning. The resulting tagging structure can be seen as a replacement to a semantic taxonomy and is assumed to be a facilitator for navigation and browsing. To give an example: The vocabulary of a personomy might contain the tag `car`. A typical categorizer (as for example depicted in Figure 1) would try to stick to the same tag instead of introducing new synonym tags such as `automobile` or `vehicle` in other contexts.

3.2 Using Tags to Describe Resources

Users who are motivated by description (so-called *describers*) aim to describe the resources they annotate accurately and precisely. As a result, their tag vocabulary typically contains an open set of tags which is dynamic by nature. Tags are not viewed as an investment into a tag structure, and changing the structure continuously is not regarded as costly. Because the tags of describers focus on describing the content of resources, these tags can be assumed to better support the process of searching and retrieval. The tag vocabulary of describers typically contains a lot of infrequently used tags and lots of synonyms (e.g. tags like `car`, `automobile` and `vehicle`). In addition, the vocabulary of a describer is likely to be larger than that of a categorizer who has mostly a stable, individual vocabulary. An example of a typical describer is depicted in Figure 2.

3.3 Discussion

While the same tag in one case might be used as a category, in another it might represent a descriptive label. So the distinction is based on a distinction with regard to the pragmatics of tagging (why and how users tag) - as opposed to the semantics of tags (what tags mean). While the distinction introduced above is theoretic, we would expect that

Authors	Categories of Tagging Motivation	Detection	Evidence	Reasoning	Systems investigated	# of users	Resources per user
Coates 2005 [3]	Categorization, description	Expert judgment	Anecdotal	Inductive	Weblog	1	N/A
Hammond et al. 2005 [5]	Self/self, self/others, others/self, others/others	Expert judgment	Observation	Inductive	9 different tagging systems	N/A	N/A
Golder et al. 2006 [4]	What it is about, what it is, who owns it, refining categories, identifying qualities, self reference, task organizing	Expert judgment	Dataset	Inductive	Delicious	229	300 (average)
Marlow et al. 2006 [14]	Organizational, social, [and refinements]	Expert judgment	N/A	Deductive	Flickr	10 (25,000)	100 (minimum)
Xu et al. 2006 [21]	Content-based, context-based, attribute-based, subjective, organizational	Expert judgment	N/A	Deductive	N/A	N/A	N/A
Sen et al. 2006 [18]	Self-expression, organizing, learning, finding, decision support	Expert judgment	Prior experience	Deductive	MovieLens	635 (3,366)	N/A
Wash et al. 2007 [20]	Later retrieval, sharing, social recognition, [and others]	Expert judgment	Interviews (semistruct.)	Inductive	Delicious	12	950 (average)
Ames et al. 2007 [1]	Self/organization, self/communication, social/organization, social/communication	Expert judgment	Interviews (in-depth)	Inductive	Flickr, ZoneTag	13	N/A
Heckner et al. 2009 [6]	Personal information management, resource sharing	Expert judgment	Survey (M. Turk)	Deductive	Flickr, Youtube, Delicious, Connotea	142	20 and 5 (minimum)
Nov et al. 2009 [16]	enjoyment, commitment, self development, reputation	Expert judgment	Survey (e-mail)	Deductive	Flickr (PRO users only)	422	2,848.5 (average)
Strohmaier et al. 2009 [19]	Categorization, description	Automatic	Simulation	Deductive	7 different datasets	2277	1,267.53 (average)

Table 1: Overview of Research on Users' Motivation for Tagging in Social Tagging Systems

users in the real world would likely be driven by a combination of both motivations, for example following a description approach to annotating most resources, while at the same time maintaining a few categories. Table 2 gives an overview of different intuitions about the two types of tagging motivation.

	Categorizer	**Describer**
Goal	later browsing	later retrieval
Change of vocabulary	costly	cheap
Size of vocabulary	limited	open
Tags	subjective	objective
Tag reuse	frequent	rare
Tag purpose	mimicking taxonomy	descriptive labels

Table 2: Intuitions about Categorizers and Describers

4. MEASURES FOR TAGGING MOTIVATION

In the following measures which capture properties of the two types of tagging motivation (Table 2) are introduced.

4.1 Terminology

Folksonomies are usually represented by tripartite graphs with hyper edges. Such graphs hold 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. [15]). Subsequently a personomy of a user $u \in U$ is the reduction of a folksonomy F to the user u ([8]). 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 $r \in R$.

4.2 Tag/Resource Ratio (trr)

Tag/resource ratio relates the vocabulary size of a user to the total number of resources annotated by this user. Describers, who use a variety of different tags for their resources, can be expected to score higher values for this measure than categorizers, who use fewer tags. Due to the limited vocabulary, a categorizer would likely achieve a lower score on this measure than a describer who employs a theoretically unlimited vocabulary. Equation 1 shows the formula used for this calculation where R_u represents the resources which were annotated by a user u. What this measure does not reflect on is the average number of assigned tags per post.

$$trr(u) = \frac{|T_u|}{|R_u|} \tag{1}$$

4.3 Orphaned Tag Ratio

To capture tag reuse, the *orphan tag ratio* of users characterizes the degree to which users produce *orphaned tags*. Orphaned tags are tags that are assigned to few resources only, and therefore are used infrequently. The *orphaned tag ratio* captures the percentage of items in a user's vocabulary that represent such orphaned tags. In equation 2 T_u^o denotes the set of orphaned tags in a user's tag vocabulary T_u based on a threshold n. The threshold n is derived from each user's individual tagging style in which t_{max} denotes the tag that was used the most. $|R_u(t)|$ denotes the number of resources which are tagged with tag t by user u. The measure ranges from 0 to 1 where a value of 1 identifies users who use orphaned tags frequently and 0 identifies users who maintain a more consistent vocabulary. Considering the categorizer - describer paradigm this would mean that categorizers would

Figure 1: Tag cloud example of a categorizer. Frequency among tags is balanced, a potential indicator for using the tag set as an aid for navigation.

Figure 2: Tag cloud example of a describer. Some tags are used often while many others are rarely used - a distribution that can be expected when users tag in a descriptive, ad-hoc manner.

be expected to be represented by values closer to 0 because orphaned tags would introduce noise to their personal taxonomy. For a describer's tag vocabulary, it would be represented by values closer to 1 due to the fact that describers tag resources in a verbose and descriptive way, and do not mind the introduction of orphaned tags to their vocabulary.

$$orphan(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t||R(t)| \leq n\}, n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \tag{2}$$

4.4 Conditional Tag Entropy (cte)

For categorizers, useful tags should be maximally discriminative with regard to the resources they are assigned to. This would allow categorizers to effectively use tags for navigation and browsing. This observation can be exploited to develop a measure for tagging motivation when viewing tagging as an encoding process, where entropy can be considered a measure of the suitability of tags for this task. A categorizer would have a strong incentive to maintain high tag entropy (or information value) in her tag cloud. In other words, a categorizer would want the tag-frequency as equally distributed as possible in order for her to be useful as a navigational aid. Otherwise, tags would be of little use in *browsing*. A describer on the other hand would have little interest in maintaining high tag entropy as tags are not used for navigation at all.

In order to measure the suitability of tags to navigate resources, we develop an entropy-based measure for tagging motivation, using the set of tags and the set of resources as random variables to calculate conditional entropy. If a user employs tags to encode resources, the conditional entropy should reflect the effectiveness of this encoding process:

$$H(R|T) = -\sum_{r \in R} \sum_{t \in T} p(r,t) log_2(p(r|t)) \tag{3}$$

The joint probability $p(r,t)$ depends on the distribution of tags over the resources. The conditional entropy can be interpreted as the uncertainty of the resource that remains given a tag. The conditional entropy is measured in bits and is influenced by the number of resources and the tag vocabulary size. To account for individual differences in users, we propose a normalization of the conditional entropy so that only the encoding quality remains. As a factor of normalization we can calculate the conditional entropy $H_{opt}(R|T)$ of an ideal categorizer, and relate it to the actual conditional entropy of the user at hand. Calculating $H_{opt}(R|T)$ can be accomplished by modifying $p(r,t)$ in a way that reflects a situation where all tags are equally discriminative while at the same time keeping the average number of tags per resource the same as in the user's personomy.

Based on this, we can define a measure for tagging motivation by calculating the difference between the observed conditional entropy and the conditional entropy of an ideal categorizer put in relation to the conditional entropy of the ideal categorizer:

$$cte = \frac{H(R|T) - H_{opt}(R|T)}{H_{opt}(R|T)} \tag{4}$$

4.5 Overlap Factor

When users assign more than one tag per resource on average, it is possible that they produce an overlap (i.e. intersection with regard to the resource sets of corresponding tags). The *overlap factor* allows to measure this phenomenon by relating the number of all resources to the total number of tag assignments of a user and is defined as follows:

$$overlap = 1 - \frac{|R_u|}{|TAS_u|} \tag{5}$$

We can speculate that categorizers would be interested in keeping this overlap relatively low in order to be able to produce *discriminative* categories, i.e. categories that are

free from intersections. On the other hand, describers would not care about a possibly high overlap factor since they do not use tags for navigation but instead aim to best support later retrieval.

4.6 Tag/Title Intersection Ratio (ttr)

In order to address the objectiveness or subjectiveness of tags, we introduce the *tag/title intersection ratio* which is an indicator how likely users choose tags from the words of a resource's title (e.g. the title of a web page). This measure is calculated by taking the intersection of the tags and the resource's title words of a specific user. At first, all resource titles occurring in a personomy are tokenized to build the set of title words TW_u. Then we filtered the tags and title words using the stop-word list which is packaged with the Snowball[1] stemmer. For normalization purposes we relate the resulting absolute intersection size to the cardinality of the set of title words.

$$ttr = \frac{|T_u \cap TW_u|}{|TW_u|} \tag{6}$$

4.7 Properties of the Presented Measures

When examining the five presented measures, we can observe that the measures focus on tagging behavior of users as opposed to the semantics of tags. This makes the introduced measures independent of particular languages. An advantage of this is that the approach is not influenced by special characters, internet slang or user specific words (e.g. "`to_read`"). In addition, the measures evaluate statistical properties of a single user personomy only; therefore knowledge of the complete folksonomy is not required.

5. EXPERIMENTAL SETUP

5.1 Dataset

For our experiments we used a dataset from Del.icio.us which is part of a larger collection of tagging datasets which was crawled from May to June 2009.[2] The requirements for the resulting datasets were the following:

- The datasets should capture complete personomies. Therefore all public resources and tags of a crawled user must be contained.
- Each post should be stored in chronological order which allows to capture changes in the tagging behavior of a user over time.
- Users who abandoned their accounts with only a few posts should be eliminated. Thus, a lower bound for the post count (R_{min}) was introduced which in the case of the Del.icio.us dataset is $R_{min} = 1000$.

The crawled Del.icio.us dataset consists of 896 users who in total used 184,746 tags to annotate 1,966,269 resources.

5.2 Correlation between Measures

Figure 3 shows the pairwise Spearman rank correlation of the proposed measures calculated on all 896 users of the Del.icio.us dataset. An interesting observation in this context is that although all measures are based on different intuitions about the motivation for tagging, some of them correlate to a great extent empirically. The two measures exhibiting highest correlation are *Tag/Resource Ratio* and *Tag/Title Intersection Ratio*, where the first measure is derived from the number of unique tags and resources and the second measure is derived from the content of the tags. Additionally these two measures also have a relatively high correlation with the other three measures. The remaining measures appear to form two separate groups. The *Orphaned Tags* and *Conditional Tag Entropy* represent one group of highly correlated measures whereas the *Overlap Factor* represents the other one. It is expected that measures with a high correlation will also show similar behavior in the evaluations.

[1] http://snowball.tartarus.org/

[2] Details of the datasets can be found in [12]

Resource	Tags User A	Tags User B
URL 1	Tag 1_A,...,Tag n_A	Tag 1_B,...,Tag m_B
⋮	⋮	⋮

Table 3: Resource Alignment - This allows human subjects to compare tagging behavior of two users w.r.t. the same resource.

Posts User A - Tag 1		Posts User B - Tag 1	
resources	tags	resources	tags
URL 1_A	$t1_A$,...,tn_A	URL 1_B	$t1_B$,...,tm_B
⋮	⋮	⋮	⋮
	⋮		⋮
Posts User A - Tag n		Posts User B - Tag n	
	⋮		⋮

Table 4: Tag Alignment - This allows human subjects to compare tagging behavior of two users w.r.t. the same tag.

6. QUALITATIVE EVALUATION

In order to assess the usefulness of the introduced measures for tagging motivation, we relate each measure to different dimensions of human judgement. Based on a subset of posts taken from users' personomies, participants of a human subject study were given the task to classify whether a given personomy represents the tagging record of a user who follows a categorization or a description approach to tagging.

To perform this task, participants were given random pairs of Del.icio.us users, for which they had to decide whether a user is a categorizer or describer. The information available to the human subjects for this task is depicted in Table 3 and 4.

6.1 Sampling

We assume that each measure is capable of making a distinction between categorizers and describers by producing high scores for describers and low scores for categorizers. For each of the five measures listed in section 4, we randomly drew five user pairs from the Del.icio.us dataset out of the measure's top 25% and bottom 25% users, reflecting

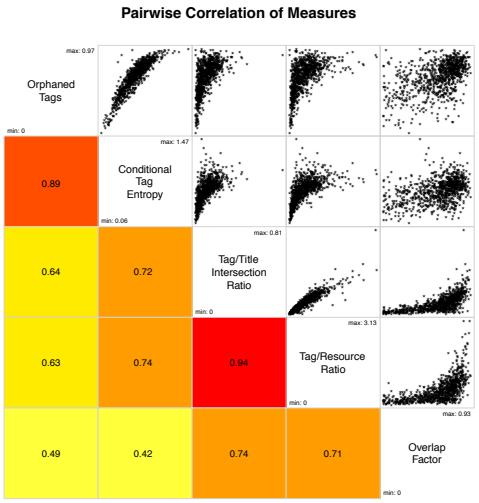


Figure 3: Spearman rank correlation of the measures for the Del.icio.us dataset (correlation values in the lower left and pairwise distribution in the upper left). All measured were developed based on different intuitions and capture different aspects of the describers and categorizers behavior, still most of them demonstrate a high agreement in regard to which users can be classified as categorizers or describers.

the set of potential categorizers and describers according to each measure. Users were chosen randomly, allowing a pair of users to be drawn from either of the two groups or from the same group.

Additionally, we ensured that the resulting user pairs are close to evenly distributed among their possible origins (top-top, top-bottom, bottom-top, bottom-bottom) to avoid a bias towards any of the two groups within our sample. With regard to the resource and tag alignment explained above, all resulting pairs had to fulfill the requirement of at least 25 shared resources and tags.

6.1.1 *Setup*

Before starting the evaluation, all participants were instructed about categorization and description as motivations for tagging in social tagging systems based on table 2. They were further provided with illustrative examples of at least two different user pairs to get used to the actual task.
Participants were then presented with 25 user pairs (one at a time) resulting from the data sampling. To simplify the task for our subjects, the resource alignment part has been restricted to a random sample of 15 shared resources while for the tag alignment part, we randomly took 5 shared tags and showed at most 5 posts for each of them. Based on this subsets of the users' personomies, the participants were instructed to perform the evaluation task.

6.2 Participants

There were three male and three female participants from an academic backgrounds with an average age of 28.5 years. Four out of six stated to have some tagging experience, one subject reported much experience, another one had low experience. According to their self-assessment, five participants characterized themselves as potential categorizers while one would characterize himself as a potential describer.

6.3 Results

6.3.1 *Inter-rater Agreement*

We calculated the inter-rater agreement for all 6 participants using both, Fleiss' Kappa as well as pairwise Cohen's Kappa which is listed in table 5. The mean pairwise Cohen's Kappa and the Fleiss' Kappa are both $\kappa = 0.51$ which can be interpreted as moderate agreement ($0.41 \leq \kappa \leq 0.60$) according to the inter-rater agreement levels of Landis and Koch (cf. [13]). The resulting kappa values appear sufficient given that our evaluation task can be considered - to some extent - subjective and complex. Participants have to de-

	P2	P3	P4	P5	P6
P1	0.40	0.43	0.72	0.44	0.56
P2		0.56	0.44	0.32	0.60
P3			0.49	0.45	0.62
P4				0.56	0.68
P5					0.40

Table 5: Pairwise Cohen's Kappa of the inter-rater agreement among 6 participants

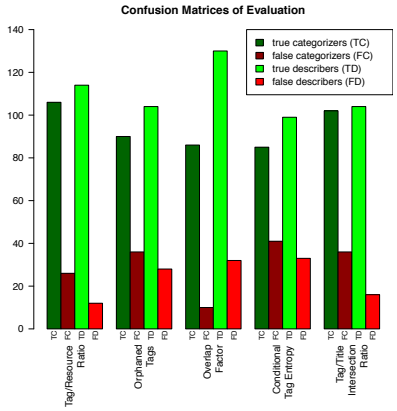


Figure 4: Confusion matrices of the evaluation

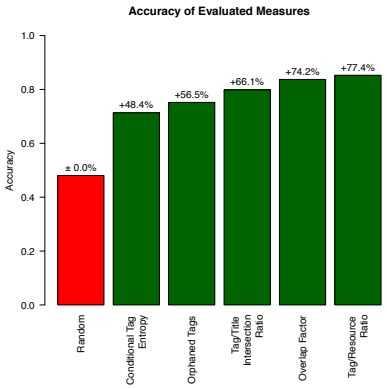


Figure 5: Accuracy for the different measures resulting from the user study

cide on a relatively small subset of a user's personomy which sometimes makes it hard to recognize the underlying motivation for tagging. Such cases may have produced subjective outcomes.

6.3.2 *Confusion Matrices*

To assess which of the five measures performs best in relation to the 6 participants' ratings for 50 users from the Del.icio.us dataset, we calculated separate confusion matrices (visually presented in Figure 4), taking each measure's classification as a potential ground truth. In subsequent analysis, all classification ratings which ended in a draw have been removed in order to get a better picture of the results achieved by every user in our study.

Figure 5 depicts the accuracy values of all measures in comparison to the random baseline, which were calculated using

$$accuracy = \frac{\#TC + \#TD}{\#TC + \#FC + \#TD + \#FD} \quad (7)$$

where TC...True Categorizer, TD...True Describer, FC...False Categorizer and FD...False Describer. The three best performing measures that achieved an accuracy of at least 0.8 are *Tag/Resource Ratio*, *Overlap Factor* and *Tag/Title Intersection Ratio*. The lowest accuracy values are held by the *Orphaned Tag Ratio* and *Conditional Tag Entropy* measures respectively.

7. QUANTITATIVE EVALUATION

In addition to qualitative evaluation, we conducted quantitative evaluation a) to assess whether the distinction between categorizers and describers has an observable impact *during tagging* and b) to evaluate which of the proposed measures best captures this distinction. Our evaluation design is based on the observation that tag recommenders influence the decisions that users make in the process of tagging (cf. [17]). We use this observation to study whether a user is influenced in his tagging decisions by different motivations for tagging: We assume that a user who is motivated by categorization would prefer a tag recommendation algorithm that suggests tags (categories) that users have used before. On the other hand, we assume that a user who is motivated by description would prefer a tag recommendation algorithm that suggests tags that are most descriptive for the resource she is tagging. Then, the extent to which one of these recommendation strategies can explain actual user behavior would be indicative of the latent influence a user is exposed to during tagging.

In our evaluation, the actual tags assigned by the user to the resource serve as ground truth. To assess the quality of the recommendation we limited the number of suggestions to a maximum of 100 tags. The set of assigned tags and recommended tags were compared and the mean average precision (MAP) over all resources and users was accumulated. In our scenario MAP is defined based on the $Precision(t)$, which is the proportion of correct tags in relation to the number of recommended tags at the rank of t:

$$MAP = \frac{1}{|U|} \sum_{u \in U} \frac{1}{|R_u|} \sum_{r \in R_u} \frac{1}{|T_{u,r}|} \sum_{t \in T_{u,r}} Precision(t) \quad (8)$$

7.1 Folksonomy-based Recommender

Given a single resource, the folksonomy-based recommender collects all tags assigned to this resource in the folksonomy. The rank of the tags is determined by their frequency. Thus, if a tag is frequently used for a specific resource, this tag will then be suggested by the folksonomy-based recommender. The folksonomy-based recommender operates on a subset of the folksonomy which is spanned only by the describers F_{desc} according to the measure being evaluated.

7.2 Personomy-based Recommender

The personomy-based recommender is based on the personal tagging vocabulary of a user. In a first step, this recommender calculates similarity of the resource to be tagged with all other resources already tagged by the user. In order

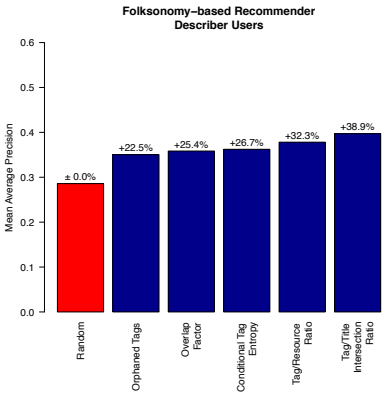


Figure 6: MAP for the set of describers influenced by the folksonomy-based recommender. All differently defined groups of describers are influenced by the folksonomy-based recommender.

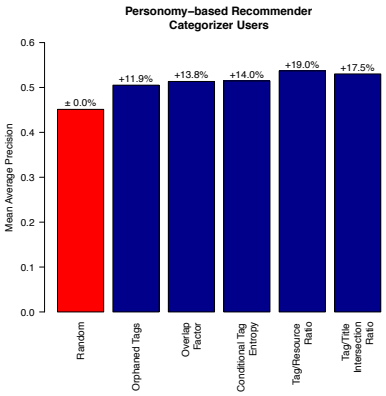


Figure 7: MAP for the set of categorizers influenced by the personomy-based recommender. All differently defined groups of categorizers are influenced by the personomy-based recommender.

to calculate the similarity of two resources, the tags of the describers within the folksonomy are exploited. The cosine similarity of the tags from the describer folksonomy F_{desc} of the two resources is taken as a proxy for the similarity of two resources. Based on the similarities of the resources, the tags from the personomy are weighted and finally ranked. Thus tags which are assigned to many resources with a high similarity value will be suggested by the personomy-based recommender.

The main goal of these recommendation strategies is not to present a novel or improved tag recommender approach, but to study the latent influence of tagging motivation on the tagging process by adopting algorithms that reflect our intuitions about why users tag. A real-world tag recommender system would have components like spam detection, tag co-occurrence statistics and others, which are not necessary for our purpose.

7.3 Tag Recommender Evaluation

To measure the effectiveness of each of the measures we compare them to a random baseline, where a user is randomly assigned to either the set of categorizers or the set of describers, building two groups of equal size. For all other measures, the users are evenly split between the two groups. The personomy-based recommender was used for categorizers, whereas the folksonomy-based recommender was used for describers. All calculations were conducted on the Del.icio.us dataset. For each measure, two sets (448 describers and 448 categorizers) were generated.

Figure 6 aims to provide an answer the question: Which user group exhibits the strongest influence from a folksonomy-based recommender? It depicts the MAP values for the different measures together with the random baseline for the describer / folksonomy-based recommender configuration. All sets of describers (as identified by the different measures) are more influenced by the folksonomy-based recommender than a random baseline group. The set of describers identified by the *Tag/Title Intersection Ratio* exhibits the strongest influence (38.9% over the baseline). We can observe the smallest influence on the set of describers identified by *Orphaned Tags*.

For the categorizer/personomy-based recommender configuration (cf. Figure 7), again all sets of categorizers are more influenced by a personomy-based recommender than a random baseline user group. Differences between differently defined groups are less pronounced compared to the folksonomy-based recommender configuration. The set of categorizers identified by the *Tag/Resource Ratio* exhibits the strongest influence (19% over the baseline). Again, the smallest influence can be observed on the set of categorizers identified by *Orphaned Tags* ratio (11.9%).

An observation that can be made is the absolute difference between the two recommender types. The recommender that is based on the personomy achieves a higher MAP for all groups of users as well as for the baseline.

The results of the evaluation reveal a latent influence on tagging behavior: Tags used by describers tend to be more similar to other describers' tags while categorizers prefer their own tagging vocabulary. Our results show that most measures capture the corresponding intuitions, but the measures *Tag/Title Intersection Ratio* and *Tag/Resource Ratio* best predict user behavior. From Figure 9 we can see that users who prefer a personomy-based recommendation algorithm can best be identified via a low *Tag/Resource Ratio*. In other words, the fewer tags a user assigns to a resource, the more likely it is that she is motivated by categorizing resources. This indicates that categorizers tend to use few tags for categorization purposes. From Figure 8 we can see that users who prefer folksonomy-based recommendations can best be identified via a high *Tag/Title Intersection Ratio*. While this result seems intuitive (describers focus on describing resources), the *Tag/Title Intersection Ratio* can

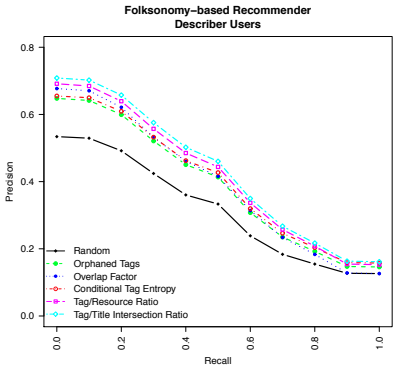


Figure 8: Precision/Recall curve for the describer users in combination with the folksonomy-based recommender. Across different recall levels, the folksonomy-based recommender influences all differently defined describer groups.

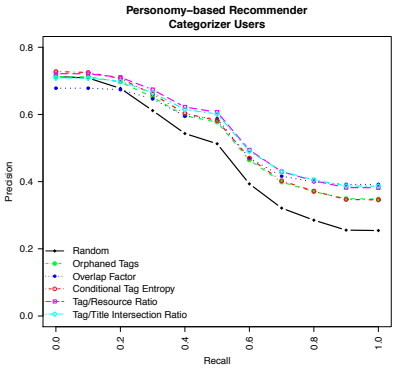


Figure 9: Precision/Recall for the categorizer/personomy-based recommender configuration. Across different recall levels, the personomy-based recommender influences all differently defined categorizer groups.

only be used on resources where title information is available (e.g. URLs). However, the results of our correlation analysis hint towards alternative, more general measures, that might be a useful approximation of the distinction between categorizers and describers on resources where title information might not be available (e.g. the *Tag/Resource Ratio*, cf. Figure 3).

8. CONCLUSION AND OUTLOOK

In this paper, we evaluated the usefulness of different measures to discriminate between categorizers and describers in social tagging systems to make the (latent) motivation behind tagging amenable to quantitative analysis. The measures introduced in this work focus on quantifying different aspects of user behavior in order to infer knowledge about a user's motivations. Knowledge about the motivation behind tagging has been found to be important for explaining folksonomical phenomena, such as the emergence of semantic structures in social tagging systems [11] or the degree to which users agree on tags [10]. The results of our qualitative evaluation show that while all measures are - to some extent - capable of approximating tagging motivation, not all are equally useful. A key finding is that the *Tag/Resource Ratio* appears to best capture human judgement. This suggests that the motivation behind tagging can - in principal - be validly approximated and integrated in folksonomical analysis with simple statistical measures. The results from our quantitative evaluation, using recommender algorithms to simulate latent influence, show that the motivation behind tagging has a significant effect on tagging behavior. The evaluation results presented in this work contribute to a deeper understanding of tagging motivation and illuminate a path towards more sophisticated approaches for studying its latent influence on the properties of tags and resulting folksonomies. While this influence has received only little attention in past research, our work represents a stepping stone for more thoroughly exploring the folksonomical effects of tagging motivation for a number of problems related to tagging systems including: 1) Search: How does the motivation behind tagging influence the performance of current folksonomy search algorithms (such as [9])? 2) Recommendation: How can current recommender algorithms explicitly consider tagging motivation to improve recommendation? and 3) Knowledge acquisition: To what extent are existing algorithms for acquiring semantic relations from folksonomies effected by tagging motivation (cf. for example [11])?

9. ACKNOWLEDGMENTS

The research presented in this work is in part funded by the Know-Center and the FWF Austrian Science Fund Grant P20269 *TransAgere*. The Know-Center is funded within

the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

10. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 971–980, New York, NY, USA, 2007. ACM.

[2] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network Properties of Folksonomies. *AI Communications*, 20:245–262, 2007.

[3] T. Coates. Two cultures of fauxonomies collide. *http://www. plasticbag. org/archives/2005/06/two_cultures_of_fauxonomies_collide/. Last access: May*, 8:2008, 2005.

[4] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198, 2006.

[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. In *SSM '08: Proceeding of the 2008 ACM workshop on Search in Social Media*, pages 3–10, New York, NY, USA, 2008. ACM.

[8] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, pages 87–102, 2006.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. FolkRank: A Ranking Algorithm for Folksonomies. In *Proc. FGIR 2006*, 2006.

[10] C. Körner. Understanding the motivation behind tagging. ACM Student Research Competition - Hypertext 2009, July 2009.

[11] C. Körner, D. Benz, A. Hotho, M. Strohmaier, and G. Stumme. Stop thinking, start tagging: Tag semantics arise from collaborative verbosity. *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, 2010. To appear.

[12] C. Körner and M. Strohmaier. A call for social tagging datasets. *ACM SIGWEB Newsletter*, 2010.

[13] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[14] 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.

[15] P. Mika. Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *The Semantic Web - ISWC 2005, Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10*, volume 3729 of *Lecture Notes in Computer Science*, pages 522–536. Springer, 2005.

[16] O. Nov, M. Naaman, and C. Ye. Motivational, Structural and Tenure Factors that Impact Online Community Photo Sharing. In *ICWSM '09: Proceedings of AAAI International Conference on Weblogs and Social Media*, 2009.

[17] 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.

[18] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer Supported Cooperative Work*, pages 181–190, New York, NY, USA, 2006. ACM.

[19] 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 23-26*, 2010.

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

[21] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland, May 2006.

3.7. Stop Thinking, Start Tagging: Tag Semantics Emerge from Collaborative Verbosity

In previous sections we have seen that there exists two different types of tagging motivation and have shown how to distinguish users by introducing statistical measures that inspect tagging behavior. An aspect which was not investigated is how these micro effects from the user propagate to the macro scale - the complete folksonomy. This can help to determine user groups that help specific automated tasks of knowledge extraction from tagging systems and reduce extensive computations.

In this paper we analyze the influence of tagging behavior on the emergent semantics within social tagging systems. Our hypothesis is that groups of users with particular usage patterns contribute more to the semantics in a folksonomy than others who are contributing "semantic noise". Using a snapshot of Delicious users we find that users who are more verbose and therefore use lots of tags (describers) are better suited for automated generation of semantic structures from a folksonomic system. Furthermore, we discovered that a subset of users (approximately 40%) is enough to produce results which are comparable or even outperform semantic precision of the complete dataset. The results of this paper point to a potential link between users' pragmatics and emergent semantics in social tagging systems.

Stop Thinking, Start Tagging: Tag Semantics Emerge from Collaborative Verbosity

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

Dominik Benz*
Knowledge & Data Engineering Group
University of Kassel
Kassel, Germany
benz@cs.uni-kassel.de

Andreas Hotho
Data Mining and Information Retrieval Group
University of Würzburg
Würzburg, Germany
hotho@informatik.uni-wuerzburg.de

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

Gerd Stumme
Knowledge & Data Engineering Group
University of Kassel
Kassel, Germany
stumme@cs.uni-kassel.de

ABSTRACT

Recent research provides evidence for the presence of emergent semantics in collaborative tagging systems. While several methods have been proposed, little is known about the factors that influence the evolution of semantic structures in these systems. A natural hypothesis is that the quality of the emergent semantics depends on the pragmatics of tagging: Users with certain usage patterns might contribute more to the resulting semantics than others. In this work, we propose several measures which enable a *pragmatic* differentiation of taggers by their degree of contribution to emerging semantic structures. We distinguish between *categorizers*, who typically use a small set of tags as a replacement for hierarchical classification schemes, and *describers*, who are annotating resources with a wealth of freely associated, descriptive keywords. To study our hypothesis, we apply semantic similarity measures to 64 different partitions of a real-world and large-scale folksonomy containing different ratios of categorizers and describers. Our results not only show that 'verbose' taggers are most useful for the emergence of tag semantics, but also that a subset containing only 40 % of the most 'verbose' taggers can produce results that match and even outperform the semantic precision obtained from the whole dataset. Moreover, the results suggest that there exists a causal link between the pragmatics of tagging and resulting emergent semantics. This work is relevant for designers and analysts of tagging systems interested (i) in fostering the semantic development of their platforms, (ii) in identifying users introducing "semantic noise", and (iii) in learning ontologies.

Categories and Subject Descriptors: H.1.2 [Information Systems]: Models and Principles [Human information processing] H.1.m [Information Systems]: Models and Principles H.3.5 [Information Storage and Retrieval]: Online Information Services[Web-based services] H.5.3 [Information Interfaces and Presentation]:Group and Organization Interfaces[Collaborative computing, Web-based interaction]

General Terms: Algorithms, Human Factors

Keywords: folksonomies, tagging, user characteristics, semantics, pragmatics

*Both authors contributed equally to this work.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

1. INTRODUCTION

Folksonomies are the core data structure of collaborative tagging systems. They are large-scale bodies of lightweight annotations provided by their user communities. Clearly, every user is following his own terminology and is only willing to a very small extent (if at all) to follow any naming conventions. Nevertheless, there is evidence for the presence of emergent semantics in such collaborative tagging systems, mainly based on tags and the folksonomical relationships between them [8, 39]. While several methods have achieved promising results for capturing emergent semantics in folksonomies (e. g., [7, 26, 36, 33, 19]), little is known about the factors that influence the evolution of semantics in these systems.

A natural hypothesis is that emergent *semantics* in folksonomies are influenced by the *pragmatics* of tagging, i. e., the tagging practices of individuals: users with certain usage patterns (cf. [14]) might contribute more to the resulting semantics than others. For example: one may assume that users who follow an 'ontology-engineering style' of tagging — i. e., users who try to maintain a "clean vocabulary" with no redundancy – contribute more to the structure of a folksonomy, which is blurred by other users who are not following this approach. However, we will show in this paper that this is *not* the case.

To this end, we will distinguish between two types of users in a folksonomy, called *categorizers* and *describers*, following the approach in [34]. Categorizers typically use a well-defined set of tags as a replacement for hierarchical classification schemes, while describers are annotating resources with a wealth of freely associated, descriptive keywords. We use a number of measures focused on capturing tagging pragmatics and approximating the membership of a user to either of the two types. These *pragmatic* measures will be used to partition a tagging dataset into subsets on which we

apply *semantic* measures [7] in order to study potential effects of tagging pragmatics on tag semantics.

Our results not only show that particular users contribute more to emerging semantics than others, but also that the "collaborative verbosity" of a fraction of *describers* can achieve and even outperform semantic precision levels obtained from the entire dataset. In summary, our results suggest that a key factor for users to be effective contributors to aggregated semantic structures is their tagging verbosity. In addition, our work provides first empirical evidence that the emergent semantics of tags in folksonomies are influenced by the pragmatics of tagging, i. e., the tagging practices of individual users.

The results of this work are relevant for researchers who want to analyze folksonomies for ontology learning purposes. For example, users who introduce "semantic noise" and hinder the semantic evolution can be identified and excluded from the data based on pragmatic measures that capture individual tagging styles of users. The proposed methods can also be used to improve and inform the design of ontology learning algorithms.

The paper is organized as follows: In section 2 we provide an overview about folksonomies and their emergent tag semantics. Section 3 deals with measures aimed at capturing different aspects of tagging pragmatics. This is followed by section 4 covering the semantic implications of tagging pragmatics in which we describe the conducted experiments and present a discussion of our results. Subsequently we give an overview of the related work (section 5). We discuss our results in the context of ontology learning and related tasks in section 6, where we also point to future work.

2. EMERGENT TAG SEMANTICS

Since the advent of folksonomies as a part of the "Web 2.0" paradigm, large corpora of human-annotated content have attracted the interest of researchers from different disciplines. In particular, there has been the early idea to study the semantics of folksonomies, e. g., work by Mika [30] or Golder and Huberman [14]. Later, more and more approaches arose to "harvest" the semantics of a folksonomy (see the section on related work for details). In many of these approaches, distributional measures were used to infer semantic relations among tags. However, in most cases the choice of these measures was done on a rather ad-hoc basis without a deeper knowledge of the semantic characteristics of each measure. A first systematic analysis which *kind* of semantic relations are returned by different measures was done by us in [7, 26]. The semantic grounding procedure presented there confirms the assumption that distributional tag relatedness measures are an appropriate means to capture the emerging semantic structures between tags in folksonomies. As our presented analysis makes strongly use of this work, we recall it here in greater detail.

2.1 Folksonomy model

In the following we will use the definition of folksonomy provided in [21]:

Definition A *folksonomy* is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, respectively. Y is a ternary relation between them, i. e., $Y \subseteq U \times T \times R$. The elements $y \in Y$ are called *tag assignments* (TAS). A *post* is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a non-empty set $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$.

Furthermore, we denote the (tag) *vocabulary* of a user as $T_u := \{t \in T \mid \exists r : (u, t, r) \in Y\}$. This represents the set of distinct tags a user has used at least once. Analogously we define $R_u := \{r \in R \mid \exists t : (u, t, r) \in Y\}$ as the set of resources a given user has tagged.

2.2 Semantic grounding of tag relatedness measures

As stated above, the notion of tag relatedness is a crucial aspect of emerging semantics in folksonomies. One way of defining it is to map the tags to a thesaurus or lexicon like Roget's thesaurus[1] or WordNet [12],[2] and to measure relatedness by means of existing semantic measures. Another option is to define measures of relatedness directly on the network structure of the folksonomy. A reason why distributional measures in folksonomies are used in addition to mapping tags to a thesaurus is the observation that the vocabulary of folksonomies often includes community-specific terms that are not included in lexical resources.

In our previous work [7] we identified several possibilities to measure tag relatedness directly in a folksonomy. Most of them use statistical information about different types of *co-occurrence* between tags, resources and users. Other approaches adopt the *distributional hypothesis* [13, 17], which states that words found in similar contexts tend to be semantically similar.

More specifically we have analyzed five measures for the relatedness of tags: the *co-occurrence count*, three context measures which capture distributional information by computing the *cosine similarity* [32] in the vector spaces spanned by users, tags, and resources, and *FolkRank* [21], a graph-based measure that is an adaptation of PageRank to folksonomies.

We observed in our experiments in [7] that the tag and resource context measures performed best, by comparing them to thesaurus-based measures based on WordNet. This indicates that the distributional hypothesis [13, 17] does not only influence the human judgment of semantic similarity [29], but also folksonomy-based distributional measures. To provide a semantic grounding of our folksonomy-based measures, we mapped the tags of a large-scale del.icio.us dataset to synsets of WordNet and used the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, we measured the similarity by using a similarity measure (JCN from here on) by Jiang and Conrath [23] that has been validated in previous user studies and applications [5].

We discovered that the context measure based on cosine similarity in a vector space that is spanned by the tags yielded an almost optimal performance at an acceptable level of computational complexity. This distributional measure is defined as follows.

The Tag Context Similarity (*TagCont*) is computed in the vector space $\mathbb{R}^T$, where, for tag t, the entries of the vector $\vec{v}_t \in \mathbb{R}^T$ are defined by $v_{tt'} := w(t, t')$ for $t \neq t' \in T$, where the weight w is the co-occurrence count , and $v_{tt} = 0$. The reason for giving weight zero between a node and itself is that we want two tags to be considered related when they occur in a similar context, and not when they occur together. TagCont is determined by using the cosine measure, a measure customary in Information Retrieval [32]: If two tags t_1 and t_2 are represented by $\vec{v}_1, \vec{v}_2 \in \mathbb{R}^T$, their cosine similarity is defined as: $\text{cossim}(t_1, t_2) := \cos \measuredangle(\vec{v}_1, \vec{v}_2) = \frac{\vec{v}_1 \cdot \vec{v}_2}{||\vec{v}_1||_2 \cdot ||\vec{v}_2||_2}$. The cosine similarity is thus independent of the length of the vectors. As in our case the vectors contain only positive entries, its value ranges from 0 (for totally orthogonal vectors) to 1 (for vectors pointing into the same direction).

By studying the taxonomic path lengths in WordNet and the number of up and down edges on the paths, we further observed that pairs of tags which had been determined as closest pairs ac-

[1] http://www.gutenberg.org/etext/22
[2] http://wordnet.princeton.edu

cording to the cosine measure and which had a path distance of 2 in WordNet were significantly more frequently siblings[3] in WordNet than pairs determined with other measures. This implied that even if the cosine measure was not able to provide an immediate synonym, it still often provided a similar tag which was on an equal level of abstraction.

In [26] we have studied further measures of tag relatedness. We discovered there that mutual information gain is yielding even more precise results. However, the quadratic complexitymakes a frequent application to numerous large-scale folksonomy subsets (as needed in our case) infeasible. Given that TagCont has been proven to make meaningful judgements of semantic tag relatedness (as shown in [7]), we use it in the remainder of this paper as a measure for emergent tag semantics.

To complement the presented semantic measures, the next section will introduce and discuss measures aimed at capturing pragmatic aspects of tagging.

3. PRAGMATICS OF TAGGING

In addition to research on emergent semantics in folksonomies, the research community has developed an interest in usage patterns of tagging, such as why and how users tag. Early work by for example Golder and Huberman [14], and later Marlow et al [27], has identified different usage patterns among users. Further work provides evidence that different tagging systems afford different tag usage and motivations [18, 16]. More recent work shows that even within the same tagging system, motivation for tagging between individual users varies greatly [34]. These observations have led to the formulation of the hypothesis that the *emergent properties of tags in tagging systems — and their usefulness for different tasks — are influenced by pragmatic aspects of tagging* [18]. If this was the case, different tagging practices and motivations would effect the processes that yield emergent semantics. This would mean that in order to assess the usefulness of methods for harvesting semantics from folksonomies, we would need to know whether these methods produce similar results across different user populations characterized by different tagging practices and driven by different motivations for tagging. Given these implications, it is interesting to explore *whether and how emergent semantics of tags are influenced by the pragmatics of tagging*.

3.1 Tagging motivation

Previous work such as [27, 16] and [18] suggests 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. On the other hand, users who are motivated by description (so called *describers*) view tagging as a means to accurately and precisely describe resources. These users tag because they want to produce annotations that are useful for later searching and retrieval. Developing a personal, consistent ontology to navigate to their resources is not their goal. Table 1 gives an overview of characteristics of the two different types of users, based on [34]. While these two types make an ideal distinction, tagging in the real world is likely to be motivated by a combination of both. A user might maintain a few categories while pursuing a description approach for the majority of resources and vice versa, or additional categories might be introduced over time. Second,

[3]An example for this are the tags '`java`' and '`python`'.

Table 1: Two Types of Taggers

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

the distinction between categorizers and describers is a distinction based on the pragmatics of tagging, and not related to tag semantics. One implication of that is that it would be perfectly plausible for the same tag (for example "`java`") to be used by both describers and categorizers, and serve both functions at the same time — for different users. In other words, the same tag might be used as a category or a descriptive label. Thereby tagging pragmatics represent an additional perspective on folksonomical data, and yet it can be expected to have effects on the emergent semantics of tags. For example, it is reasonable to assume that the tags produced by describers are more descriptive than tags produced by categorizers. If this was the case, algorithms focused on utilizing tags for ontology learning would benefit from knowledge about the users' motivation for tagging.

3.2 Measures of tagging pragmatics

Because the motivation behind tagging is difficult to measure without direct interaction with users, we use this distinction as an inspiration for the definition of the following surrogate measures for pragmatic aspects of tagging only.

3.2.1 Vocabulary size

$$vocab(u) = |T_u| \tag{1}$$

The *vocabulary size* (as proposed by for example [14] or [27]) reflects the number of tags found in a user's tag vocabulary T_u. Describers would likely produce an open set of tags with a unlimited and dynamic tag vocabulary while categorizers would try to keep their vocabulary limited and would need far fewer tags. A deficit of this measure is that it does not reflect on the total number of annotated resources, which are considered in the next measure.

3.2.2 Tag/resource ratio (trr)

$$trr(u) = \frac{|T_u|}{|R_u|} \tag{2}$$

This measure relates the vocabulary size with the total number of annotated resources. Taggers who use lots of different tags for their resources would score higher values for this measure than users that use fewer tags. Due to the limited vocabulary, a categorizer would likely achieve a lower score on this measure than a describer who employs a theoretically unlimited vocabulary. The equation above shows the formula used for this calculation where R_u represents the resources which were annotated by a user u. What this measure does not reflect on is the average number of assigned tags per post. This is considered next.

3.2.3 Average tags per post (tpp)

$$tpp(u) = \frac{\sum^{r} |T_{ur}|}{|R_u|} \tag{3}$$

This measure quantifies how many tags a user applies to a resource on average. Taggers who usually apply lots of tags to their re-

sources get higher scores by this measure than users who use few tags during the annotation process. Describers would score high values for this measure because of their need for detailed and verbose tagging. In contrast categorizers would score lower values because they try to annotate their resources in an efficient way.

3.2.4 Orphan ratio

$$orphan(u) = \frac{|T_u^o|}{|T_u|}, T_u^o = \{t||R(t)| \leq n\}, n = \left\lceil \frac{|R(t_{max})|}{100} \right\rceil \quad (4)$$

As a final measure, we introduce the *orphan ratio* of users to capture the degree to which users produce *orphaned tags*. Orphaned tags are tags that users assign to just a few resources. The *orphan ratio* thus captures the percentage of items in a user's vocabulary that represent such orphaned tags. T_u^o denotes the set of orphaned tags in a user's tag vocabulary T_u (based on a threshold n). The threshold n is derived from each user's individual tagging style in which t_{max} denotes the tag that was used the most. $|R(t)|$ denotes the number of resources which are tagged with tag t by user u. The measure ranges from 0 to 1 where a value of 1 identifies users with lots of orphaned tags and 0 identifies users who maintain a more consistent vocabulary. Considering the categorizer - describer paradigm this would mean that categorizers tend more towards values of 0 because orphaned tags would introduce noise to their personal taxonomy. For a describer's tag vocabulary, this measure would produce values closer to 1 due to the fact that describers tag resources in a verbose and descriptive way, and do not mind the introduction of orphaned tags to their vocabulary.

3.3 Properties of measures

While these measures of tagging pragmatics were inspired by the dichotomy between categorizers and describers, we do not require them to accurately capture this distinction. Another aspect is that these measures might not only capture intrinsic user characteristics, but can also be influenced by e.g. elements of user interfaces (such as recommenders). What is important in the light of our hypothesis is that all of *the above measures are independent of semantics* — they capture *usage patterns* of tagging (the pragmatics of tagging) only. This allows us to explore a potential link between tagging pragmatics and the emergent semantics of tags.

4. SEMANTIC IMPLICATIONS OF TAGGING PRAGMATICS

As detailed in Sec. 2.2, the distributional hypothesis states that words used in similar contexts tend to have similar meanings. As tags in a folksonomy can be regarded as natural language entities, a crucial question is how to identify an adequate context for capturing their semantics. However, given the massive amounts of data available in social tagging systems, the question is not only to identify a *valid* context, but also to identify the *minimal* context which retains the relevant structures while allowing for efficient computation. As human annotators are the creators of implicit semantic structures, an important aspect hereby is which users should be included in an optimal context composition. Following our discussion in the prior section, our hypothesis is that individual tagging pragmatics can play an important role for selecting "productive" users. The question is whether the categorizers — who follow the ontology engineering principle of a clean vocabulary — or the describers — who provide more descriptions to their resources — are the more "productive" ones.

In order to answer this question, our strategy is to analyze the suitability of each of our previously introduced pragmatic measures to assemble a (preferentially small) subset of users which provides a sufficient context to harvest emergent tag semantics. The general idea hereby is to start at both ends of the scale with the "extreme" categorizers and describers, and then to subsequently add more users (in the order given by the respective measure). In each step, we check how well the folksonomy partition defined by the current user subset serves as a basis to compute semantically related tags. For the latter, we revert to the tag context relatedness measure that has shown to produce valid results (cf. Sec. 2). The assumption hereby is that this TagCont measure will yield more closely related tags when better implicit semantic structures are present. Hence, this whole procedure allows us to assess the quality of the emergent semantics and finally the degree to which tagging pragmatics have influenced its evolution.

Table 2: del.icio.us dataset statistics.

dataset	$\|T\|$	$\|U\|$	$\|R\|$	$\|Y\|$
full	10,000	511,348	14,567,465	117,319,016
min100res	9,944	100,363	12,125,476	96,298,409

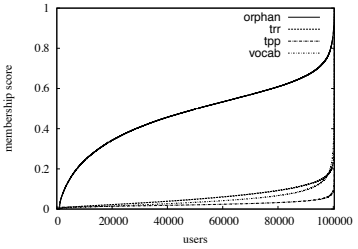


Figure 1: Distribution of the membership scores for each introduced measure of tagging motivation (orphan ratio, tag/resource ratio, tags per post and vocabulary size), computed for the 100,393 users present in our del.icio.us dataset (x-axis). Values close to 0 on the y-axis indicate strong categorizers, while values close to one 1 point to describer users. All measures were normalized to the interval $[0, 1]$.

4.1 Experiments

The goal of our experiments is to quantify the influence of individual tagging practices on emergent tag semantics in a folksonomy. We will first provide details on our dataset and then explain each experimentation step before discussing the results.

4.1.1 Description of the dataset

In order to validate our hypothesis on real-world data, we used a dataset crawled from the social bookmarking system del.icio.us in November 2006.[4] In total, data from 667,128 users of the del.icio.us community were collected, comprising 2,454,546 tags, 18,782,132 resources, and 140,333,714 tag assignments. As our experimental methodology involves the comparison with semantically related tags obtained from the full dataset, we need to ensure that the quality of those is high. Because the applied tag relatedness measure is based on the co-occurrence of tags with other tags, the inherent sparseness of infrequent tags makes them less useful for our purpose. Hence, we stick to our dataset containing the 10,000 most

[4]All data sets used in this study are publicly available at http://www.kde.cs.uni-kassel.de/benz/papers/2010/www.html

frequent tags of del.icio.us, and to the resources/users that have been associated with at least one of those tags. We will refer to the resulting folksonomy as the *full* dataset (see Table 2).

In order to eliminate noise introduced by our measures misjudging new users, we furthermore removed all users having less than 100 resources in their collection. The reason behind this is that e. g., the tag/resource ratio is not very informative in the case of a new user with very few resources. Interestingly, our result shows that removing this "long tail" of new (or inactive) users already increases the quality of the learned semantic relations. Details of this observation will be discussed in Section 4.3. We will denote the resulting dataset as *min100res* (see Table 2).

4.1.2 Experimental setup

In order to assess the capability of each of our measures to predict "productive" users, we followed an incremental approach: For each of our measures $m \in \{$*orphan,vocab,trr,tpp*$\}$, we first created a list L_m of all users $u \in U$ sorted in ascending order according to $m(u)$. All our measures yield low values for categorizers, while giving high scores to describers. This means that e.g. the first user in the orphan ratio list (denoted as $L_{orphan}[1]$) is assumed to be the most extreme categorizer, while the last one ($L_{orphan}[k], k = |U|$) is assumed to be the most extreme describer. Figure 1 depicts the obtained distribution of membership scores for each ordered list L_{tpp}, L_{trr}, L_{orphan} and L_{vocab}. An observation which can be made in this figure is that the distribution of the $orphan$ measure differs clearly from the other three measures. This implies that the orphan ratio seems to be able to make more fine-grained distinction between users. However, our results did not exhibit a positive impact on the resulting semantics; rather contrary, the orphan ratio performs often worse than the other measures (see section 4.2 for details).

Because we are interested in the minimum amount of users needed to provide a valid context, we start at both ends of L and extract two folksonomy partitions CF_1^m and DF_1^m based on 1% of the "strongest" categorizers ($Cat_1^m = \{L_m[i] \mid i \leq 0.01 \cdot |U|\}$) and describers ($Desc_1^m = \{L_m[i] \mid i \geq 0.99 \cdot |U|\}$). $CF_1^m = (CU_1^m, CT_1^m, CR_1^m, CY_1^m)$ is then the sub-folksonomy of F induced by Cat_1^m, i. e., it is obtained by $CU_1^m := Cat_1^m$, $CY_1^m := \{(u,t,r) \in Y | u \in Cat_1^m\}$, $CT_1^m := \pi_2(CY_1^m)$, and $CR_1^m := \pi_3(CY_1^m)$. The sub-folksonomy DF_1^m is determined analogously.

As a next step, we took the first extracted partition CF_1^m as input to extract semantic tag relations, in the way described in Section 2.2. We check whether the data produced by a very small subset of "extreme" categorizers already suffices to compute meaningful semantic relations. More specifically, for each tag $t \in CT_1^m$, we computed its most similar tag t_{sim} according to the tag context relatedness defined in [7]. We then looked up each resulting pair (t, t_{sim}) in WordNet and measured – whenever both t and t_{sim} were present — the Jiang-Conrath distance $JCN(t, t_{sim})$ between both words (see Sec. 2.2). After that we took the average JCN distance of all mapped tag pairs as an indicator of the quality of emergent semantic structures contained in CF_1^m:

$$JCN_{avg}(CF_1^m) = \frac{\sum_{t \in CT_1^m} JCN(t, t_{sim})}{wn\_pairs(CT_1^m)}$$

Here, $wn\_pairs(DT_1^m)$ denotes the number of tag pairs (t, t_{sim}) (i. e., a tag and its most similar tag) for which both t and t_{sim} are present in WordNet. The corresponding describer partition DF_1^m was processed in the same manner.

As discussed in Sec. 2.2, we use the Jiang-Conrath distance as an indicator of the "true" semantic relatedness between tags. However, in order to avoid the dependency of our results on a single measure of semantic similarity, we also measured the *taxonomic path length* for each mapped tag pair (t, t_{sim}) between the two synsets s_1 and s_2 containing t and t_{sim}, respectively.[5] This measure counts the number of nodes in the WordNet subsumption hierarchy along the shortest path between s_1 and s_2. We noticed that the judgements of both measures (JCN and taxonomic path length) were almost perfectly correlated throughout our experimentation; for this reason, we will stick to the JCN distance in the remainder of this paper, because it has been shown to be a better surrogate for the human perception.

We repeated this overall procedure for each of our measures $m \in \{$*orphan,vocab,trr,tpp*$\}$ and for the following user fractions i:

$$i \in \{1, 2, 3, \ldots, 24, 25, 30, 40, 50, 60, 70, 80, 90\}$$

As we keep adding users while incrementing i, it is important to notice that the size of the resulting "sub-folksonomy" is growing towards the size of the full dataset i. e., $DF_{100}^m = CF_{100}^m = F$. Another important aspect is the fact that users are added in descending order of their membership degree in the respective user class: This means that CF_1^m contains users u who score high on measure m, while e. g., CF_{50}^m contains a more mixed population. "Mixed" in this context means that there exist users in CF_{50}^m which are to a certain degree assumed to exhibit describer characteristics as measured by m. This implies that the distinction between both user groups is blurred while incrementing i. In other words, one can also read these partitions from the other side, namely that CF_{90}^m contains all users *except* 10% of the most extreme describers.

So in summary, we created 64 partitions for each of our 4 measures (32 categorizer + 32 describer), summing up to a total of 256 sub-folksonomies, each being extracted by a different composition of users according to their tagging characteristics. Before presenting our results on the most suitable partitions for extracting semantic tag relations, we discuss upper and lower bounds. As we measured the quality of an extracted relation between two tags t and t_{sim} by its Jiang-Conrath distance within WordNet, a lower bound can be identified by computing the pairwise JCN distance between all tags $t \in T$ and averaging over the minimum distance found for each tag:

$$JCN_{lower}(F) = \frac{\sum_{t \in T} \min_{t_{sim} \in T} JCN(t, t_{sim})}{wn\_pairs(T)}$$

As an upper bound we assume that the respective folksonomy subset does not contain any inherent semantics and hence only randomly related tags are returned by our measure. We simulate this by defining a random relatedness function $rand(t)$, which returns a randomly selected tag $t_{sim} \in T, t_{sim} \neq t$. The upper bound is then:

$$JCN_{upper}(F) = \frac{\sum_{t \in T} JCN(t, rand(t))}{wn\_pairs(T)}$$

For the del.icio.us dataset it turned out that $JCN_{upper} \approx 15.834$ and $JCN_{lower} \approx 0.758$. Please recall that JCN is a semantic *distance* measure — which means a low JCN distance corresponds to a high degree of semantic relatedness.

As seen later (cf. Figure 2), none of our experimental conditions (including the full dataset) came close to the lower bound. There are (at least) two explanations for this. Firstly, the lower bound was determined independently of a sub-folksonomy of the full dataset. It would be interesting to determine that sub-folksonomy that provides the optimal average Jiang-Conrath distance. Then one could check how far it is away from this optimum, and one could try to

[5] If t and t_{sim} were present in more than one synset, we took the shortest possible path.

learn a classifier for this target dataset. Unfortunately, the computation of this sub-folksonomy requires the consideration of all subsets of the user set U and is thus computationally unfeasible.

Secondly, WordNet is built by language experts with the goal to capture *all* existing senses of a given word. Given two tags t_1 and t_2, our JCN implementation searched for the smallest possible distance between *any* two senses of each tag. By doing so for all possible pairs of tags $t \in T$, the probability is quite high to find two quite closely related (or even equal) senses. Contrary to that, the technophile bias of the user population of del.icio.us leads to some usage-induced relations which are not reflected well within WordNet; as an example, the most related tag to `doom` in a folksonomy subset was `quake`, leading to a large JCN distance of ≈ 18.08, while the optimal distance was found between `doom` and `will` with ≈ 1.88. This observation does not invalidate the procedure of semantic grounding as a whole, because we *do* find matching semantics in both systems. The same approach has also been taken in previous publications focused on measures for semantic relatedness [7].

4.2 Results

In Figures 2(a) and 2(b) we present the results of our analysis of the different sub-folksonomies which were created in each of our 256 experimental conditions.

The horizontal axis displays the percentage of included users; the vertical axis displays the average JCN distance obtained from computing semantically related tags based on the respective partition. The dashed line at the bottom of each figure represents the level of semantic precision obtained from the full dataset.

A first impression is — in all diagrams, independently of the selection strategy — that mass matters: the average JCN distance decreases and hence the results get better while more users are included. This equally holds for the random selection strategy (solid line, $+$). In other words, the more people contribute to a collaborative tagging system, the higher is the quality of the semantic tag relations which can be obtained from the folksonomy structure they produce. This matches the intuition that a sufficient "crowd" is necessary to facilitate the emergence of the "wisdom of the crowds".

However, the obvious differences between the two Figures 2(a) and 2(b) suggests that the composition of the crowd also seems to make a difference: When incrementally adding users ordered from categorizers to describers (starting from the left of Figure 2(a)), all resulting folksonomy partitions yield systematically weaker semantic precisions compared to adding users in random order (solid line, $+$). This effect can be observed most clearly for the vocabulary size measure *vocab* (dotted line, ▲), which judges users as categorizers when the size of their tag vocabulary is small (see Sec. 3.2.1). Only after the addition of 90 % of all users in this order, the quality of the inherent semantics are on the same level of randomly selected 90 %. The other measures — with an exception of the tags per post ratio (dotted line, •) which will be discussed later — show a very similar behavior, namely the tag/resource ratio (dotted line, ■) and the orphan ratio (dotted line, $*$).

When incrementally building sub-folksonomies starting from describer users (Figure 2(b)), we see a completely different picture: most measures start on the same or even on a slightly higher level of contained semantics compared to adding users in a random order. Beginning from roughly 10 % included users, all sub-folksonomies yield better results than the random case. In addition, after having added 40 % of the users in the order of the tag/resource ratio (dotted line, □), we can even observe a first improvement of the results compared with the full dataset. This implies that a bit less than the "better half" of the complete folksonomy population produces equally precise semantic structures compared to the whole unfiltered "crowd". This improvement increases and reaches its maximum after adding 70 % of all users, before it decreases again to the global level.

Table 3: Statistical properties of selected folksonomy partitions. %t denotes the fraction of the tags from the complete dataset included in the respective partition; %w denotes the number of similar tag pairs (t, t_{sim}) found in WordNet for the respective partition divided by the number of mapped pairs from the whole dataset. For the entire dataset, $|T| = 9944$ and *wn_pairs*$(T) = 4335$.

	DF_i^{trr}		DF_i^{tpp}		DF_i^{orphan}		DF_i^{vocab}	
i	%t	%w	%t	%w	%t	%w	%t	%w
1	0.93	1.03	0.96	1.01	0.97	1.02	0.98	1.04
3	0.96	1.02	0.98	1.02	0.99	1.01	0.99	1.03
5	0.97	1.02	0.99	1.02	0.99	1.02	0.99	1.03
10	0.97	1.03	0.99	1.02	1.00	1.02	0.99	1.01
20	0.98	1.02	0.99	1.00	1.00	1.03	0.99	1.01
50	0.98	1.02	1.00	1.00	1.00	1.00	1.00	1.01
70	0.99	1.01	1.00	1.00	1.00	1.00	1.00	1.00

	CF_i^{trr}		CF_i^{tpp}		CF_i^{orphan}		CF_i^{vocab}	
i	%t	%w	%t	%w	%t	%w	%t	%w
1	0.56	0.48	0.44	0.00	0.48	0.59	0.27	0.18
3	0.86	0.77	0.74	0.23	0.78	0.77	0.59	0.44
5	0.94	0.83	0.87	0.49	0.89	0.88	0.76	0.59
10	0.97	0.90	0.95	0.80	0.95	0.95	0.91	0.78
20	0.99	0.95	0.97	0.88	0.97	0.98	0.97	0.88
50	1.00	1.00	0.98	0.96	0.98	1.01	0.98	0.95
70	1.00	1.00	0.98	0.98	0.99	0.99	0.98	0.98

Especially for very small partitions (roughly ≤ 20 %), users selected in descending order by their vocabulary size yield the best results (dotted line, △). Interestingly, this effect is inverse when adding users the other way round (dotted line, ▲, in Fig. 2(a)): Even quite a large number of users with small vocabularies perform considerably worse than most other folksonomy partitions. This means that scale still matters, as the quality almost constantly increases while adding users; but the "collaborative verbosity" of a small subset of users with large vocabularies seems to lead to much richer inherent semantics than the contributions of a larger set of more "tight-lipped" users.

One could suspect now that this comparison is not completely fair: Especially when selecting users with small vocabularies, the question is to which extent semantic relations *can* be present at all in the data. In other words: If the aggregated small vocabularies of a subset of categorizers result in a considerably smaller global vocabulary compared to aggregating more verbose users, then the probability to find semantically close tags would consequently be much lower. In the worst case, the vocabulary would be so small that the "right partner" for a given tag *does not exist*.

In order to eliminate this concern, we counted the size of the collective tag vocabulary for each sub-folksonomy. In addition, we measured how many tag pairs (t, t_{sim}) could be mapped to WordNet during the computation of the JCN distance. By doing this we want to make sure that the average semantic distance is computed roughly over the same number of tag pairs. Table 3 summarizes some selected statistics relative to the complete dataset.[6]

The first observation is that in all partitions based on describers (upper half of the table) the global vocabulary is almost completely contained (≥ 93 %). For partitions larger than 20 %, this value raises to 98 %. The same holds for the fraction of tag pairs mapped

[6] We did not include the statistics for every partition for space reasons; missing values can be interpolated from the given examples.

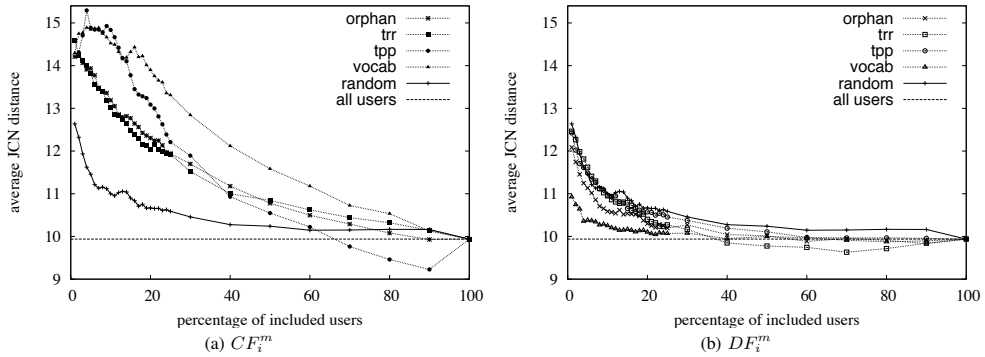


Figure 2: Average Jiang-Conrath distance between pairs of semantically related tags computed from different folksonomy partitions. The partitions were created based on user subsets as determined by different pragmatic measures (orphan ratio, tag/resource ratio, tags per post, vocabulary size). Each datapoint corresponds to a "sub-folksonomy" CF_i^m (a) / DF_i^m (b) with $i = 1, 2, \ldots, 25, 30, 40, 90$ (from left to right in both cases). The x-axis denotes the percentage of all folksonomy users included in the subset, and the y-axis depicts the quality of the semantic tag relations obtained from the respective partition by means of the JCN distance. In Figure 2(a), users were added ordered from categorizers to describers, and in Figure 2(b) ordered in the reverse direction. (Note: Empirical lower-/upper bounds are $\approx 0.758/15.83$, respectively; cf. Sec. 4.1.2.)

to WordNet. On the first sight, values > 1 might appear counter-intuitive here. The explanation is the following: It can happen that for a given tag t that its most similar tag t_{sim} based on the complete dataset it not present in WordNet, but its most similar tag t'_{sim} based on a particular partition is contained. A high percentage of mapped tags does not imply better semantics per se (as the two mapped tags can still be semantically distant); but the comparison of different sub-folksonomies is more meaningful when they both allow for a roughly equal number of mapped pairs. As expected, the coverage observed for the describer-based case is not as complete for the categorizer-based excerpt: For very small samples, the collective tag pool is in fact small. However, this effect is mitigated already for samples of 3 %; and starting from roughly 10-20 % sample size, a sufficient global vocabulary exists ($\approx$ 97 %). This means that the comparison in general is performed on a fair basis, because the underlying vocabulary sizes are comparable.

Our results suggest that sub-folksonomies based on describers contain more precise inherent semantic structures than partitions based on categorizers. However, there seems to be a limitation with this observation: Inspecting the curve for the *tpp* measure on the right side of Figure 2(a), one can observe that the most precise semantic relations among all experimental conditions are found after the addition of 90 % of the categorizers according to this measure. As stated above, this partition can also be read from the other side and corresponds to a removal of 10 % of the most extreme describers. As the *tpp* measure captures the average numbers of tags per post, there seems to be a number of "ultra-taggers" who use a large number of tags per post (many spammers, typically more than 9 tags per post in our case) have detrimental effects on the global tag semantics. In other words, removing these users seems to eliminate "semantic noise", leading to more precise tag semantics.

4.3 Discussion and implications

Recent research demonstrated that the collective output of tagging systems can be used for harvesting emergent semantic structures from the web [35, 33, 7]. Our results show that the effectiveness of current semantic measures for tag relatedness are influenced by factors originating outside of the semantic realm. On small data samples (up to 40 % of users in our dataset), we have singled out a group of users (categorizers) that has particularly detrimental effects on the performance of current semantic measures compared to random sampling. At the same time, describers (based on the tags-per-resource measure) consistently outperform random sampling, and can level and even outperform the results achieved on the entire dataset with as little as 40 % of users. This suggests that methods for harvesting *semantics* from samples of tagging systems can be made more effective when utilizing knowledge about the *pragmatics* of tagging, considering individual user behavior. For analysts of small data samples who wish to improve semantic relatedness measures, this would mean focusing on those users that use tagging systems in a verbose 'Stop Thinking, Start Tagging' fashion. With increasing sample sizes (>50 % of users), we can observe that adding more categorizers does not produce significantly better results. However, when adding more describers, we see significant improvements in performance until we hit an accuracy limit at approximately 90 % of users. This suggests that rewarding verbose taggers comes with limitations itself: The most verbose taggers (in our case: mostly spammers) negatively influence the results as well.

The practical implications of our results concern mainly two questions: (i) What is the minimum amount of users needed to produce meaningful tag semantics in collaborative tagging systems and how can these users be selected? (ii) Does the quality of emerging tag semantics increase with the available amount of data, or can it be improved by eliminating "semantic noise"?

A main contribution of our analysis lies in the observation that tagging pragmatics, i. e., individual tagging characteristics, play an important role in both cases. The experiments described above reveal that not all users contribute equally to emerging semantics; we could show that a relatively small subset of describers yields significantly better results than a group of categorizers. Figure 3

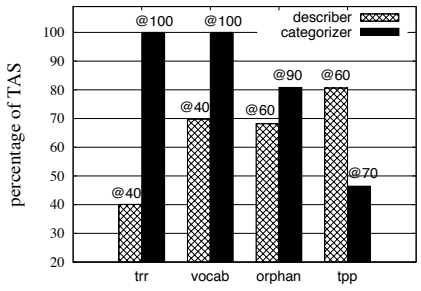


Figure 3: Minimum size of the folksonomy partitions created by each measure sufficient to reach the semantic precision of the complete dataset. The y-axis denotes the percentage of tag assignments contained in the smallest folksonomy partition which reached the global semantic precision; the labels above the bars depict the percentage of users the respective sub-folksonomies are based on.

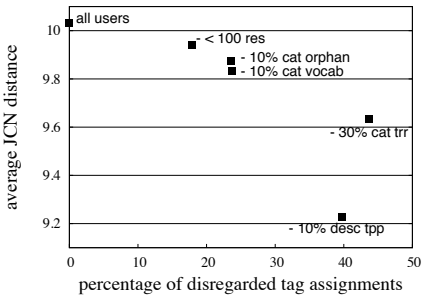


Figure 4: Improvement of semantic precision by removing users from the complete dataset. The y-axis depicts the semantic precision of the (sub-)folksonomies, while the x-axis denotes the percentage of tag assignments which were disregarded by removing certain users. The label at each data point describes which users were removed.

summarizes the minimum sizes of the folksonomy partitions identified by each of our introduced measures necessary to reach the level of semantic precision for the entire dataset. The white bars correspond to sampling users ordered from describers to categorizers (Fig. 2(b)) while the black bars correspond to sampling users ordered in the opposite direction (Fig. 2(a)). The number on top of each bar displays the user fraction needed to reach the global semantic precision; the y-axis depicts the size of the respective sub-folksonomy relative to the complete one.

In general, most describer-based selection strategies create smaller folksonomies which produce meaningful semantics. The "smallest" one consists of 40 % describers according to the trr measure, responsible for roughly 40 % of all tag assignments. However, the observation that uncontrolled verbosity is not a good thing is confirmed by the fact that removing 30 % of the most extreme describers according to the tags-per-post measure (rightmost black bar) also creates a comparatively small and semantically precise partition. According to Figure 3, two adequate strategies for creating the smallest possible scaffolding for global tag semantics can be identified: (1) include roughly half of the users with a high tag/resource ratio, and (2) remove roughly one third of "ultra-taggers" identified by a large average number of tags per post.

The next interesting question to ask is whether, and to which extent we can even infer *more precise* semantics when removing users. Figure 4 displays the obtained semantic precision (y-axis) plotted against the amount of tag assignments removed when removing users according to different selection strategies. The first and most simple strategy is to remove the "long tail" of users with less than 100 resources in their collection. This already eliminates roughly 18 % of the data, while interestingly slightly improving the semantic precision. One cannot conclude from that that the long tail of users does not contain valuable information at all. But with regard to *popular* tags (recall that we restricted our dataset to the top 10.000 tags), a valid first insight is that the long tail of inactive users can be discarded during the computation of semantic tag relations.

As discussed before, our results indicate that categorizers also have a detrimental effect on the quality of the emerging structures. Removing 30 % of them as determined by the tag/resource ratio leads to a further improvement in semantic precision. The best result in all of our experimental conditions however was reached by eliminating 10 % of the extreme describers according to the tags-per-post measure. Those "hyper-active" users (in our case mostly spammers as confirmed by manual inspection) generate roughly 40 % of the global amount of tag assignments. Spammers typically use a large number of semantically disjoint tags to attract other users and to bias search engines towards their posted URLs. Unsurprisingly, they are not very helpful for creating meaningful tag relations. Rather the contrary is the case: we can see in our results that spammers introduce significant semantic noise — a removal of them leads to an overall improvement in accuracy of the resulting semantic structures. Turning the tables around, this insight can of course also be useful for spammer detection itself — but because our dataset does not contain explicit spammer labels for each user, determining the exact ratio of spammers detected by each of our pragmatic measures is subject to future work.

4.4 Generalization on other datasets

In order to exclude the possibility that the implications mentioned above are influenced by characteristics from the del.icio.us dataset, we repeated the experimental procedure described in section 4.1.2 on a dataset from January 2010 of our own social bookmarking system BibSonomy[7]. It contained 17,777 users, 10,000 tags and 4,520,212 resources connected by 34,505,061 TAS. Space does not permit a detailed presentation of the results; but in general, all measures exhibited a very similar behavior as observed for the del.icio.us dataset in Figures 2(a) and 2(b). Especially the practical implications discussed in Section 4.3 were valid in a nearly identical way for the BibSonomy data: 30% of describers according to the trr measure were sufficient to reach the semantic precision of the whole dataset, and removing 20% of describers according to the tpp measure led to the best overall semantics.

5. RELATED WORK

There is series of research discussing folksonomies from a formal [30] and informal [28] perspective. First quantitative analysis of folksonomies are provided in [14] and the underlying structure is

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

analyzed in [8]. Tag-based metrics for resource distance have been introduced in [6]. [1] gives evidence that social annotations are a potential source for generating semantic metadata.

Many publications on folksonomies introduce measures for tag relatedness, e. g., [19, 33]. However, the choice of a specific measure of relatedness is often made without justification and often it appears to be rather ad hoc. Which context information captures the meaning of tags best has been addressed by [38]. Questions that have not been addressed previously include which users contribute to what extent to emergent semantics in folksonomies, and to what extent are tag semantics influenced by tagging pragmatics. In [7] we performed first analysis on different kinds of relatedness measures and different types of semantic relationships. In the paper at hand, we investigate different measures to characterize users and their level of contribution to the semantics of a folksonomy. To the best of our knowledge, no other analysis in the literature addresses the interrelation between pragmatic aspects of tagging (namely user characteristics) and their semantic implications for tag relatedness.

[25] generalizes standard tree-based measures of semantic similarity to the case where documents are classified in the nodes of an ontology with non-hierarchical components. The measures introduced there were validated by means of a user study. [31] analyses distributional measures of word relatedness and compares them with measures of semantic relatedness in thesauri like WordNet. In [26] we provide a systematic analysis of a broad range of similarity measures that can be applied directly and symmetrically to build networks of users, tags, or resources and to compute similarities between these entities.

A task which depends heavily on quantifying tag relatedness is that of tag recommendation in folksonomies. In the last years, a lot of research activities can be observed as two ECML PKDD discovery challenges [20, 11] were based on this topic. Existing work in general can be broadly divided in approaches that analyze the content of the tagged resources with information retrieval techniques [4] and approaches that use collaborative filtering methods based on the folksonomy structure [37]. An example of the latter class of approaches is [22]. Relatedness measures also play a role in assisting users who browse the contents of a folksonomy. [3] shows that navigation in a folksonomy can be enhanced by suggesting tag relations grounded in content-based features.

A considerable number of investigations are motivated by the vision of "bridging the gap" between the Semantic Web and Web 2.0 by means of ontology-learning procedures based on folksonomy annotations. [30] provides a model of semantic-social networks for extracting lightweight ontologies from del.icio.us. Other approaches for learning taxonomic relations from tags are provided by [19, 33]. Another branch of research is concerned with the enrichment of folksonomies by including data from existing semantic repositories and ontologies [2]. [24] proposes an RDFS model to formalize the meaning of tags relative to other tags. [15] presents a generative model for folksonomies and also addresses the learning of taxonomic relations. [39] applies statistical methods to infer global semantics from a folksonomy. The results of our paper are especially relevant to inform the design of such learning methods.

6. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the influence of individual tagging practices in collaborative tagging systems on the emergence of global tag semantics. After proposing a number of statistical measures to assign users to two broad classes of categorizers and describers, we systematically built folksonomy partitions by incrementally adding users from each class. We then judged the quality of the emergent semantics contained in each of these "subfolksonomies" by means of semantically grounded tag relatedness measures. Apart from the observation that adding more users is beneficial in many — but not all — cases, our results reveal a dependence of the obtained semantic structures on the different partitions. In general, the collaborative verbosity of describers provides a better basis for harvesting meaningful tag semantics. However, this observation comes with a limitation: The most verbose taggers (in our case mostly spammers) negatively influenced semantic accuracy. From a practical perspective, the pragmatic measures can be used to select a comparatively small subset of users which produce tag relations of equal or better quality than the entire set of users. In addition, the measures can facilitate improvement of the global semantic precision by eliminating users that introduce "semantic noise". Experiments with an additional dataset corroborate the assumption that our findings can be generalized to other collaborative tagging systems.

A main implication of our work is the presentation of first empirical evidence for a causal link between the pragmatics of tagging (individual tagging practices) and the emergent semantics of tags. This link is *not* dependent on our choice for a particular semantic relatedness measure, because 1) the chosen Jiang-Conrath distance has been shown to best reflect human judgements of semantic relatedness in previous validation studies [5] and 2) our experiments with alternative measures for semantic relatedness have produced similar results (cf. section 4.1).

This finding has a number of interesting implications for related areas of research: 1) While our results focus on semantic relatedness, it appears plausible that other semantic tasks, such as hypo/hypernym detection, exhibit similar effects. We argue that a general link between tagging pragmatics and tag semantics could yield new ways of thinking and new algorithm designs for learning ontologies from folksonomies. 2) Current tag recommender algorithms tap into semantic relations between tags in order to recommend tags to users. Our results suggest that knowledge about why and how users tag could help to further improve the performance of tag recommender systems. 3) Utilizing tag recommenders to influence tagging behavior and to direct the evolution of folksonomies towards more precise emergent semantics seems to represent an exciting and promising area for future work.

7. ACKNOWLEDGMENTS

The research presented in this work is in part funded by the Know-Center, the FWF Austrian Science Fund Grant P20269, the WebZubi project funded by BMBF and the VENUS project funded by Land Hessen.

8. REFERENCES

[1] H. S. Al-Khalifa and H. C. Davis. Exploring the value of folksonomies for creating semantic metadata. 2007.

[2] S. Angeletou. Semantic enrichment of folksonomy tagspaces. *The Semantic Web - ISWC 2008*, pages 889–894, 2009.

[3] M. Aurnhammer, P. Hanappe, and L. Steels. Integrating collaborative tagging and emergent semantics for image retrieval. In *Proc. WWW2006, Collaborative Web Tagging Workshop*, May 2006.

[4] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW06: Proc. of the 15th Int'l Conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006.

[5] A. Budanitsky and G. Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[6] C. Cattuto, A. Baldassarri, V. D. P. Servedio, and V. Loreto. Emergent community structure in social tagging systems. *Advances in Complex Physics*, 2007. Proc. of the European Confeence on Complex Systems ECCS2007.

[7] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *The Semantic Web – ISWC 2008, Proc.Intl. Semantic Web Conference 2008*, volume 5318 of *LNAI*, pages 615–631, Heidelberg, 2008. Springer.

[8] C. Cattuto, C. Schmitz, A. Baldassarri, V. D. P. Servedio, V. Loreto, A. Hotho, M. Grahl, and G. Stumme. Network properties of folksonomies. *AI Communications Journal, Special Issue on Network Analysis in Natural Sciences and Engineering*, 20(4):245–262, 2007.

[9] D. Chandler. *Semiotics: The Basics*. Taylor & Francis, second edition, 2007.

[10] F. de Saussure. *Course in General Linguistics*. Duckworth, London, [1916] 1983. (trans. Roy Harris).

[11] F. Eisterlehner, A. Hotho, and R. Jäschke, editors. *ECML PKDD Discovery Challenge 2009 (DC09)*, volume 497 of *CEUR-WS.org*, Sept. 2009.

[12] C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

[13] J. R. Firth. A synopsis of linguistic theory 1930-55. *Studies in Linguistic Analysis (special volume of the Philological Society)*, 1952-59:1–32, 1957.

[14] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, April 2006.

[15] H. Halpin, V. Robu, and H. Shepard. The dynamics and semantics of collaborative tagging. In *Proc. of the 1st Semantic Authoring & Annotation Workshop (SAAW)*, 2006.

[16] T. Hammond, T. Hannay, B. Lund, and J. Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11(4), April 2005.

[17] Z. S. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.

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

[19] P. Heymann and H. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, CS dep., April 2006.

[20] A. Hotho, D. Benz, R. Jäschke, and B. Krause, editors. *ECML PKDD Discovery Challenge 2008 (RSDC'08)*. Workshop at 18th Europ. Conf. on Machine Learning (ECML'08) / 11th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'08), 2008.

[21] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The Semantic Web: Research and Applications*, volume 4011 of *LNAI*, pages 411–426, Heidelberg, 2006. Springer.

[22] R. Jäschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Proc. PKDD 2007*, volume 4702 of *Lecture Notes in Computer Science*, pages 506–514, Berlin, Heidelberg, 2007.

[23] J. J. Jiang and D. W. Conrath. Semantic Similarity based on Corpus Statistics and Lexical Taxonomy. In *Proc. of the International Conference on Research in Computational Linguistics (ROCLING)*. Taiwan, 1997.

[24] F. Limpens, F. Gandon, and M. Buffa. Collaborative semantic structuring of folksonomies. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:132–135, 2009.

[25] A. G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani. Algorithmic computation and approximation of semantic similarity. *World Wide Web*, 9(4):431–456, 2006.

[26] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *18th International World Wide Web Conference*, pages 641–641, April 2009.

[27] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HT'06: Proc. of the 17th conference on Hypertext and Hypermedia*, pages 31–40, New York, NY, USA, 2006.

[28] A. Mathes. Folksonomies – Cooperative Classification and Communication Through Shared Metadata, December 2004.

[29] S. Mcdonald and M. Ramscar. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proc. of the 23rd Annual Conference of the Cognitive Science Society*, pages 611–6, 2001.

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

[31] S. Mohammad and G. Hirst. Distributional measures as proxies for semantic relatedness. Submitted for publication, http://ftp.cs.toronto.edu/pub/gh/Mohammad+Hirst-2005.pdf.

[32] G. Salton. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[33] P. Schmitz. Inducing ontology from Flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, May 2006.

[34] M. Strohmaier, C. Körner, and R. Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. Technical report, Knowledge Management Institute - Graz University of Technology, 2009.

[35] H. Wu, M. Zubair, and K. Maly. Harvesting social knowledge from folksonomies. In *HYPERTEXT '06: Proc. of the seventeenth conference on Hypertext and hypermedia*, pages 111–114, New York, NY, USA, 2006. ACM Press.

[36] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *WWW '06: Proc. of the 15th international conference on World Wide Web*, pages 417–426, New York, NY, USA, 2006. ACM.

[37] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proc. of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland, May 2006.

[38] C. A. Yeung, N. Gibbins, and N. Shadbolt. Contextualising tags in collaborative tagging systems. In *HT '09: Proc. of the 20th ACM conference on Hypertext and hypermedia*, pages 251–260, New York, NY, USA, 2009. ACM.

[39] L. Zhang, X. Wu, and Y. Yu. Emergent semantics from folksonomies: A quantitative study. *Journal on Data Semantics VI*, 2006.

3.8. Tags vs Shelves - From Social Tagging to Social Classification

In the previous section we have seen that users who describe documents play a larger role in the emergence of semantics in a social tagging system than those who categorize their resources with the help of tags. This raises the question as to which tasks are best suited to exploiting data produced by categorizers rather than describers.

In this experiment we examine the usefulness of different types of tagging behavior on the task of automatically classifying documents - in this case books - into categories of the Library of Congress and the Dewey Classification Scheme. For our experiments we used snapshots of the GoodReads[22] and the LibraryThing [23] systems. The obtained results show that users who are driven by categorization outperform users who produce descriptive tags while not topping the complete data. In addition we argue that tag suggestions given by the system can have an impact on the observed behavior. This is an observation that was also made in other systems like the Mendeley system.

[22] http://www.goodreads.com/
[23] http://www.librarything.com

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, Classification, 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.

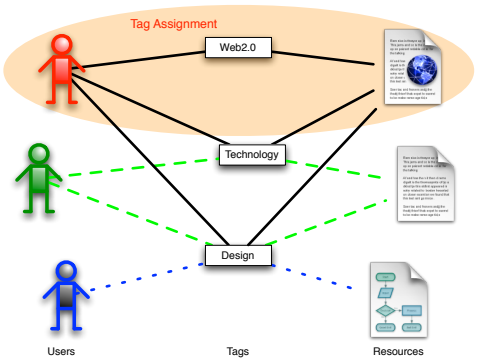


Figure 1: A folksonomy comprising users, tags and resources. A bookmark refers to all tags a user applies to a given resource.

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.

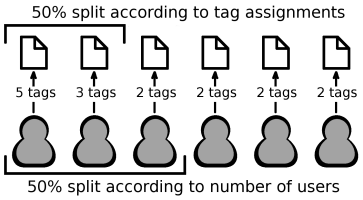


Figure 3: Example of a 50% segment by splitting based on tag assignments or number of users. Splitting by number of users is unfair, since it may yield bigger amounts of data.

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.

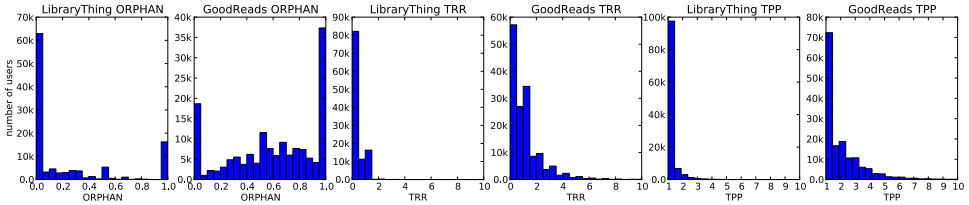


Figure 2: Histogram with distributions of the three studied measures (ORPHAN, TPP and TRR) for the two datasets. X axis represents the quantiles of values, whereas Y axis represents the number of users for each quantile.

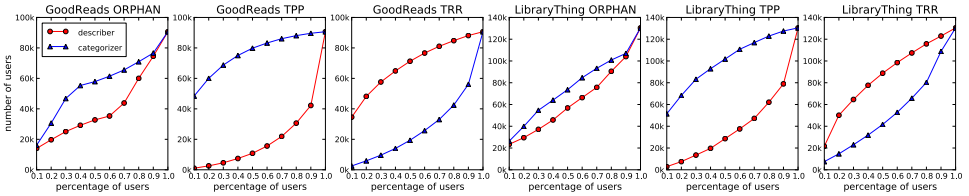


Figure 4: Number of users in each of the subsets. The X axis represents the percents of selected top users, ranging from 10% to 100% of tag assignments, with a step size of 10%, whereas the Y axis represents the number of users. Different user numbers result from the variety of user behavior which are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

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

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.

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

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-

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).

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

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).

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

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

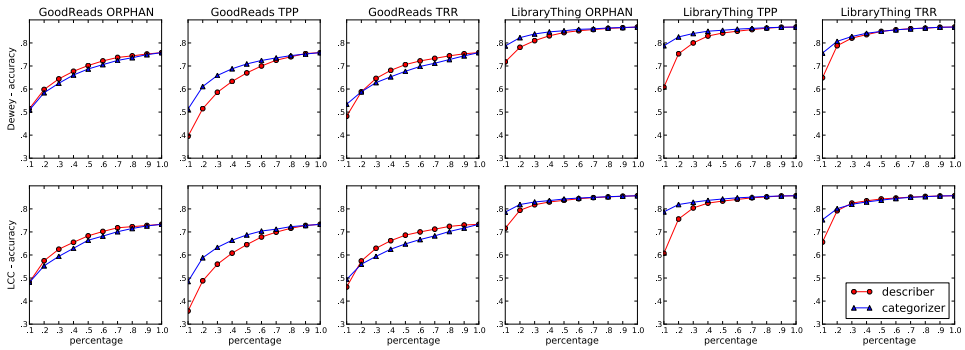


Figure 5: LCC and Dewey Accuracy of LibraryThing and GoodReads. The X axis represents the percents of selected top users, ranging from 10% to 100% and with a step size of 10%, either for Categorizers or Describers, whereas Y axis represents the accuracy. As can be seen *TPP* scores the best accuracy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

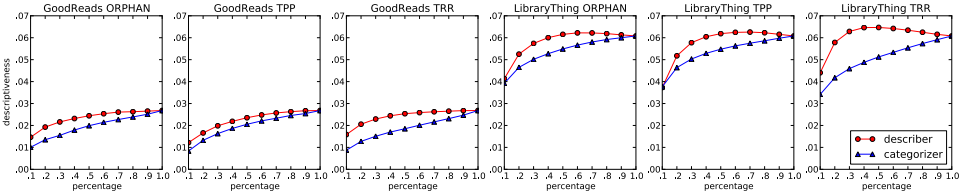


Figure 6: Descriptiveness of LibraryThing and GoodReads. The X axis represents the percents of selected top users, ranging from 10% to 100% and with a step size of 10%, either for Categorizers or Describers, whereas Y axis represents the degree of similarity to descriptive data. When splitting up the 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.

3.9. One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata

The last section is a step towards future research of analyzing tagging behavior. Measuring the abstractness levels of tags using the underlying folksonomy is a potential utility for a new differentiation between *generalists* and *specialists*. An interesting hypothesis is that users who use mainly specific words as tags have a deeper understanding of the things they annotate whereas users who are new to a topic or have little knowledge of it apply more general tags.

In this paper we evaluate different methods to measure term abstractness (in other words "tag generality") in folksonomic systems. For the grounding we use dataset snapshots from Yago[24], WordNet[25], DMOZ[26] and WikiTaxonomy[27] and compare them systematically to the earlier introduced measures. We find that measures that examine both tag entropy and tag network centrality serve well for abstractness approximation. For centrality evaluation purposes the tag co-occurrence graph seems to be a better basis for these generality evaluations than the tag similarity graphs we construct for our experiments. Another contribution of this paper is an additional perspective on the "generality vs. popularity" problem indicating that the popularity of a tag is a good gauge for its generality.

In the future these tag generality measures may be used for an expert identification task in folksonomies e.g. in combination with measures introduced in [Stirling, 2007].

[24] http://www.mpi-inf.mpg.de/yago-naga/yago/
[25] http://wordnet.princeton.edu/
[26] http://www.dmoz.org/
[27] http://www.h-its.org/english/research/nlp/download/wikitaxonomy.php

One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata

Dominik Benz*[1], Christian Körner*[2],
Andreas Hotho[3], Gerd Stumme[1], and Markus Strohmaier[2]

[1] Knowledge and Data Engineering Group (KDE), University of Kassel
{benz,stumme}@cs.uni-kassel.de
[2] Knowledge Management Institute, Graz University of Technology
{christian.koerner,markus.strohmaier}@tugraz.at
[3] Data Mining and Information Retrieval Group, University of Würzburg
hotho@informatik.uni-wuerzburg.de

Abstract. Recent research has demonstrated how the widespread adoption of collaborative tagging systems yields emergent semantics. In recent years, much has been learned about how to harvest the data produced by taggers for engineering light-weight ontologies. For example, existing measures of tag similarity and tag relatedness have proven crucial step stones for making latent semantic relations in tagging systems explicit. However, little progress has been made on other issues, such as understanding the different levels of tag generality (or tag abstractness), which is essential for, among others, identifying hierarchical relationships between concepts. In this paper we aim to address this gap. Starting from a review of linguistic definitions of word abstractness, we first use several large-scale ontologies and taxonomies as grounded measures of word generality, including Yago, Wordnet, DMOZ and WikiTaxonomy. Then, we introduce and apply several folksonomy-based methods to measure the level of generality of given tags. We evaluate these methods by comparing them with the grounded measures. Our results suggest that the generality of tags in social tagging systems can be approximated with simple measures. Our work has implications for a number of problems related to social tagging systems, including search, tag recommendation, and the acquisition of light-weight ontologies from tagging data.

Keywords: tagging, generality, measures, emergent semantics, folksonomies

1 Introduction

Since the advent of participatory web applications like Flickr[4], Youtube[5] or Delicious[6], social annotations (especially in the form of collaboratively created

* Both authors contributed equally to this work.
[4] http://www.flickr.com
[5] http://www.youtube.com
[6] http://www.delicious.com

keywords or *tags*) form an integral part of current approaches to collaborative knowledge management. Analyses of the structure of the resulting large-scale bodies of human-annotated resources have shown several interesting properties, especially regarding the presence of *emergent semantics*. Motivated by the vision of bridging the gap towards the Semantic Web, much has been learned in recent years about how to harvest the data produced by taggers for engineering light-weight ontologies. However, little progress has been made on other issues, such as understanding the different levels of tag generality (or tag abstractness), which is essential for e.g. identifying hierarchical relationships between concepts. While several methods of deriving taxonomies from tagging systems have been proposed, a systematic comparison of the underlying notion of abstractness is largely missing.

This paper aims to address this gap by presenting a systematic analysis of various folksonomy-derived notions of term abstractness. Starting from a review of linguistic definitions of word abstractness, we first use several large-scale ontologies and taxonomies as grounded measures of word generality, including Yago, Wordnet, DMOZ and WikiTaxonomy. Then, we introduce and apply several folksonomy-based methods to measure the level of generality of given tags. We evaluate these methods by comparing them with the grounded measures.

Our results show that the abstractness judgments by some of the measures under consideration come close to those of well-defined and manually built taxonomies. Furthermore, we provide empirical evidence that tag abstractness can be approximated by simple measures. The results of this research are relevant to all applications who benefit from a deeper understanding of tag semantics, e.g. ontology learning or clustering algorithms, tag recommendations systems or folksonomy navigation facilities. In addition, our results can help to alleviate the problem of varying "basic levels" in folksonomies [11] by matching more specific terms (used usually by domain experts) to more general ones.

This paper is structured as follows: At first we give an overview about related work, especially regarding term abstractness and emergent semantics (Section 2). This is followed by some basic notions in Section 3. In the subsequent section we give an overview of the introduced measures (section 4) and evaluate them in Section 5 with the help of established datasets as ground truth and a user study. Finally we conclude in Section 6 and point to future work.

2 Related Work

The first research direction relevant to this work has its roots in the analysis of the structure of collaborative tagging systems. Golder and Huberman [11] provided a first systematic analysis, mentioning among others the hypothesis of "varying basic levels" – according to which users use more specific tags in their domain of expertise. However, the authors only provided exemplary proofs for this hypothesis, lacking a well-grounded measure of tag generality. In the following, a considerable number of approaches proposed methods to make the implicit semantic structures within a folksonomy explicit [19, 13, 22, 2]. All of

the previous works comprise in a more or less explicit way methods to capture the "generality" of a tag (e.g. by investigating the centrality of tags in a similarity graph or by applying a statistical model of subsumption) – however, a comparison of the chosen methods has not been given. Henschel et al. [12] claim to generate more precise taxonomies by an entropy filter. In our own recent work [17] we showed that the quality of semantics within a social tagging system is also dependent on the tagging habits of individual users, Heymann [14] introduced another entropy-based tag generality measure in the context of tag recommendation.

From a completely different point of view, the question of which factors determine the generality or abstractness of natural language terms has been addressed by researchers coming from the areas of Linguistics and Psychology. The psychologist Paivio [20] published in 1968 a list of 925 nouns along with human concreteness rankings; an extended list was published by Clark [8]. Kammann [16] compared two definitions of word abstractness in a psychological study, namely imagery and the number of subordinate words, and concluded that both capture basically independent dimensions. Allen et al. [1] identify the generality of texts with the help of a set of "reference terms", whose generality level is known. They also showed up a correlation between a word's generality and its depth in the WordNet hierarchy. In their work they developed statistics from analysis of word frequency and the comparison to a set of reference terms. In [25], Zhang makes an attempt to distinguish the four linguistic concepts fuzziness, vagueness, generality and ambiguity.

3 Basic Notions

As stated above, the main intent of a term generality measure is to allow a differentiation of lexical entities $l_1, l_2, \ldots$ by their degree of abstractness (i.e. their ability to "bind" other tags). As a prerequisite for a formalization of this problem, we will first introduce a common terminology which allows us to refer to the usage of lexical entities in the context of taxonomies and collaborative tagging systems in a unified way.

Taxonomies, Core Ontologies and Lexicons First of all, according to [7] a *taxonomy* can also be regarded as a part of a *core ontology*, [3] $\mathbb{O} := (C, root, \geq_C, L^C, \mathcal{F})$, whereby C is a set of concept identifiers and *root* is a designated root concept for the partial order $\geq_C$ on C. $\geq_C$ is called concept hierarchy or *taxonomy*; if $c_1 \geq_C c_2 (c_1, c_2 \in C)$, then c_1 is a *superconcept* of c_2, and hence we assume c_1 to be more abstract or "general" than c_2. L^C is a set of lexical labels for concepts and a mapping relation $\mathcal{F}$ which associates concepts with their respective label. Please note that a concept c can be associated with one or more labels, i.e. $\forall l \in L^C : |\{l : (c, l) \in \mathcal{F}\}| \geq 1$. As an example, in scientific contexts the terms "article" and "paper" are often used synonymously, which would be reflected by $(c_1, paper) \in \mathcal{F}$ and $(c_1, article) \in \mathcal{F}$, given that c_1 is the concept

identifier of scientific articles. In the literature, one often defines a separate *lexicon* $\mathbb{L} = (L^C, \mathcal{F})$ and associates it with a core ontology [18]; but as it suffices for the context of this work, we assume the lexicon to be an integral part of the ontology itself for the sake of simplicity.

Folksonomies As an alternative approach to taxonomies, collaborative tagging systems have gained a considerable amount of attention. Their underlying data structure is called *folksonomy*; according to [15], a folksonomy is a tuple $\mathbb{F} := (U, T, R, Y)$ where U, T, and R are finite sets, whose elements are called *users*, *tags* and *resources*, respectively. Y is a ternary relation between them, i.e. $Y \subseteq U \times T \times R$. An element $y \in Y$ is called a *tag assignment* or TAS. A *post* is a triple (u, T_{ur}, r) with $u \in U$, $r \in R$, and a non-empty set $T_{ur} := \{t \in T \mid (u, t, r) \in Y\}$. Intrinsically, concepts are not explicitly present within a folksonomy; however, the set of tags T contains lexical items similar to the vocabulary set L^C of a core ontology.

Term Graphs Both core ontologies and folksonomies introduce various kinds of relations among the lexical items contained in them. A typical example are *tag cooccurrence networks*, which constitute an aggregation of the folksonomy structure indicating which tags have occurred together. Generally spoken, these *term graphs* $\mathbb{G}$ can be formalized as weighted undirected graphs $\mathbb{G} = (L, E, w)$ whereby L is a set of vertices (corresponding to lexical items), $E \subseteq L \times L$ model the edges and $w\colon E \to \mathbb{R}$ is a function which assigns a weight to the edges. As an example, given a folksonomy (U, T, R, Y), one can define the post-based[7] *tag-tag cooccurrence graph* as $\mathbb{G}_{cooc} = (T, E, w)$ whose set of vertices corresponds to the set T of tags. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The *weight* of this edge is given by the number of posts that contain both t_1 and t_2, i.e. $w(t_1, t_2) := \text{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}$

As we will define term abstractness measures based on core ontologies, folksonomies and term graphs, we will commonly refer to them as *term structures* $\mathbb{S}$ in the remainder of this paper. $L(\mathbb{S})$ is a projection on the set of lexical items contained in $\mathbb{S}$. Based on the above terminology, we now formally define a term abstractness measure in the following way:

Definition 1. *A* term abstractness measure $\sqsupset^{\mathbb{S}}$ *based upon a term structure* $\mathbb{S}$ *is a partial order among the lexical items* L *present in* $\mathbb{S}$*, i.e.* $\sqsupset^{\mathbb{S}} \subseteq L(\mathbb{S}) \times L(\mathbb{S})$*. If* $(l_1, l_2) \in \sqsupset^{\mathbb{S}}$ *(or* $l_1 \sqsupset^{\mathbb{S}} l_2$*) we say that* l_1 *is more abstract than* l_2*.*

In the following, we will make frequent use of *ranking functions* $r\colon L(\mathbb{S}) \to \mathbb{R}$ for lexical items in order to define a tag abstractness measure; please note that a ranking function corresponds to a partial order according to $(l_1, l_2) \in \sqsupset^{\mathbb{S}} \Leftrightarrow r(l_1) > r(l_2)$. We will denote the resulting term abstractness measure as $\sqsupset^{\mathbb{S}}_r$.

[7] Other possibilities are resource-based and user-based cooccurrence; we use post-based cooccurrence in the scope of this work as it is efficiently computable and captures a sufficient amount of information.

4 Measures of Tag Generality

Based on the notions defined above, we will now introduce a set of ranking functions r which are supposed to order lexical items within a folksonomy $\mathbb{F}$ by their degree of abstractness, inducing a partial order $\sqsupset_r^{\mathbb{F}}$ among the set of tags.[8] The measures are partially based on prior work in related areas, and build on different intuitions. One commonality they all share is that none of them considers the textual content of a tag itself (e.g. with linguistic methods). All measures operate solely on the folksonomy structure itself or on a derived term network, making them language-independent.

Frequency-based measures A first natural intuition is that more abstract tags are simply used more often, because there exist more resources which they describe – as an example, the number of "computer"s in the world is much larger than the number of "notebook"s; hence one might assume that within a folksonomy, the tag "computer" is used more often than the tag "notebook". We capture this intuition in the abstractness measure $\sqsupset_{freq(t)}^{\mathbb{F}}$ induced by the ranking function *freq* which counts the number of tag assignments according to $freq(t) = \text{card}\{(u, t', r) \in Y : t = t'\}$

Entropy-based measures Another intuition stems from information theory: Entropy measures the degree of uncertainty associated with a random variable. Considering the application of tags as a random process, one can expect that more general tags show a more even distribution, because they are probably used at a relatively constant level to annotate a broad spectrum of resources. Hence, more abstract terms will have a higher entropy. This approach was also used by Heymann [14] to capture the "generality" of tags in the context of tag recommendation. We adapt the notion from there and define

$$entr(t) = -\sum_{t' \in cooc(t)} p(t'|t) \log p(t'|t) \tag{1}$$

whereby $cooc(t)$ is the set of tags which cooccur with t, and $p(t'|t) = \frac{w(t',t)}{\sum_{t'' \in cooc(t)} w(t'',t)}$ (with $w(t', t)$ being the cooccurrence weight defined in Section 3). $entr(x)$ induces the term abstractness measure $\sqsupset_{entr}^{\mathbb{F}}$.

Centrality Measures In network theory the centrality of a node $v \in V$ in a network G is usually an indication of how important the vertex is [24]. Applied to our problem at hand, centrality can also be contemplated as a measure of abstractness or generality, following the intuition that more abstract terms are also more "important". We adopted three standard centralities (degree, closeness, betweenness). All of them can be applied to a term graph $\mathbb{G}$, leaving us

[8] Note that all term abstractness measures based on real-value ranking functions are by construction total orders, but this is not mandatory.

with three measures $\sqsupseteq^{\mathbb{G}}_{dc}$, $\sqsupseteq^{\mathbb{G}}_{bc}$ and $\sqsupseteq^{\mathbb{G}}_{cc}$ as follows: *Degree centrality* simply counts the number of direct neighbors $d(v)$ of a vertex v in a graph $G = (V, E)$:

$$dc(v) = \frac{d(v)}{|V| - 1} \tag{2}$$

According to *betweenness centrality* a vertex has a high centrality if it can be found on many shortest paths between other vertex pairs:

$$bc(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3}$$

Hereby σ_{st} denotes the number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths between s and t passing through v. As its computation is obviously very expensive, it is often approximated [4] by calculating the shortest paths only between a fraction of points. Finally, a vertex ranks higher according to *closeness centrality* the shorter its shortest path length to all other reachable nodes is:

$$cc(v) = \frac{1}{\sum_{t \in V \setminus v} d_G(v, t)} \tag{4}$$

$d_G(v, t)$ denotes hereby the geodesic distance (shortest path) between the vertices v and t.

Statistical Subsumption Schmitz et.al. [22] applied a statistical *model of subsumption* between tags when trying to infer hierarchical relationships. It is based on the assumption that a tag t subsumes another tag t' if $p(t|t') > \xi$ and $p(t'|t) < \xi$ for a suitable threshold ξ. For measuring generality, the number of subsumed tags can be seen as an indicator of abstractness – the more tags a tag subsumes the more general it is:

$$subs(t) = \text{card}\{t' \in T : p(t|t') > \xi) \wedge p(t'|t) < \xi\} \tag{5}$$

5 Evaluation

In order to assess the quality of the tag abstractness measures $\sqsupseteq^{\mathbb{F}}_{freq}$, $\sqsupseteq^{\mathbb{F}}_{entr}$, $\sqsupseteq^{\mathbb{G}}_{dc}$, $\sqsupseteq^{\mathbb{G}}_{bc}$, $\sqsupseteq^{\mathbb{G}}_{cc}$ and $\sqsupseteq^{\mathbb{F}}_{subs}$ introduced above, a natural approach is to compare them against a ground truth. A suitable grounding should yield reliable judgments about the "true" abstractness of a given lexical item. Of special interest are hereby taxonomies and concept hierarchies, whose hierarchical structure typically contains more abstract terms like "entity" or "thing" close to the taxonomy root, whereby more concrete terms are found deeper in the hierarchy. Hence, we have chosen a set of established core ontologies and taxonomies, which cover each a rather broad spectrum of topics. They vary in their degree of controlledness – WordNet (see below) on the one hand being manually crafted by language experts, while the Wikipedia category hierarchy and DMOZ on the other hand are built in a much less controlled manner by a large number of motivated web users. In the following, we first briefly introduce each dataset; an overview about their statistical properties can be found in Table 1.

Table 1: Statistical properties of the datasets used in the evaluation.

Core ontology	$\|C\|$	$\|\geq_C\|$	$\|L^C\|$	$\|\mathcal{F}\|$	$\|\sqsupset^{\mathbb{0}}\|$
WORDNET	79,690	81,866	141,391	141,692	2,028,925
YAGO	244,553	249,465	206,418	244,553	2,078,788
WIKI	2,445,974	4,447,010	2,445,974	2,445,974	13,171,439
DMOZ	767,019	767,019	241,910	767,019	5,210,226

Folksonomy	$\|U\|$	$\|T\|$	$\|R\|$	$\|Y\|$
DEL (Delicious)	667,128	2,454,546	18,782,132	140,333,714

Term Graphs	$\|T\|$	$\|E\|$
COOC	892,749	38,210,913
SIM	10,000	405,706

5.1 Grounding Datasets

WordNet [9] is a semantic lexicon of the English language. In WordNet, words are grouped into *synsets*, sets of synonyms that represent one concept. Among other relations, the *is-a* relation connects a *hyponym* (more specific synset) to a *hypernym* (more general synset). A synset can have multiple hypernyms, so that the graph is not a tree, but a directed acyclic graph. In order to allow for comparison with the other grounding datasets, we focussed on the noun subsumption network[9]. As it consists of several disconnected hierarchies, it is useful to add a fake top-level node subsuming all the roots of those hierarchies, making the graph fully connected and allowing a relative abstractness judgment between all contained pairs of nouns.

Yago [23] is a large ontology which was derived automatically from Wikipedia and WordNet. Manual evaluation studies have shown that its precision (i.e. the percentage of "correct" facts) lies around 95%. It has a much higher coverage than WordNet (see Table 1), because it also contains named entities like people, books or products. The complete ontology contains 1.7 million entities and 15 million relations; as our main interest lies in the taxonomy hierarchy, we restricted ourselves to the contained *is-a* relation[10] among concepts.

WikiTaxonomy [21] is the third dataset used for evaluation. This large scale domain independent taxonomy[11] was derived by evaluating the semantic network between Wikipedia concepts and labeling the relations as *isa* and *notisa*, using methods based on the connectivity of the network and on lexico-syntactic patterns. It contains by far the largest number of lexical items (see Table 1), but this comes at the cost of a much lower level of manual controlledness.

DMOZ[12] (also known as the open directory project or ODP) is an open content directory for links of the World Wide Web. Although it is hierarchically structured, it differs from the above-mentioned datasets insofar as its internal link structure does not always reflect a sub-concept/super-concept relationship. Despite this fact, we we included the DMOZ category hierarchy as a grounding

[9] `http://wordnet.princeton.edu/wordnet/download/` (v2.1)
[10] `http://www.mpi-inf.mpg.de/yago-naga/yago/subclassof.zip` (v2008-w40-2)
[11] `http://www.h-its.org/english/research/nlp/download/wikitaxonomy.php`
[12] `http://www.dmoz.org/`

dataset because it was built for a similar purpose like many collaborative bookmarking services (namely organizing WWW references). In addition, some of its top level categories (like "arts" or "business") are described by rather abstract terms.

5.2 Tagging Dataset

In order to test the performance of our proposed term abstractness measures, we used a dataset crawled from the social bookmarking system Delicious in November 2006.[13] From the raw data, we first derived the *tag-tag cooccurrence graph* $COOC = (T', E_{cooc}, w_{cooc})$. Two tags t_1 and t_2 are connected by an edge, iff there is at least one post (u, T_{ur}, r) with $t_1, t_2 \in T_{ur}$. The edge weight is given by $w_{cooc}(t_1, t_2) := \text{card}\{(u, r) \in U \times R \mid t_1, t_2 \in T_{ur}\}$. In order to exclude cooccurrences introduced by chance and to enable an efficient computation of the centrality measures, we removed all tags from the resulting graph with a degree of less than 2.

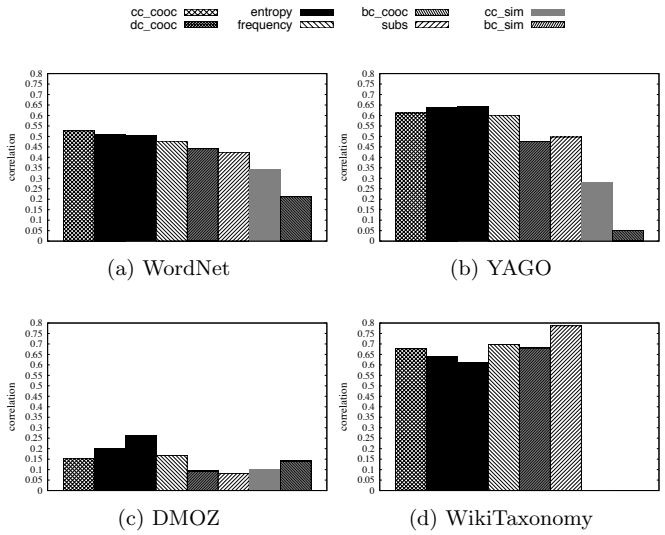
In a similar way to [13], we also derived a *tag-tag similarity graph* $SIM = (T'', E_{sim}, w_{sim})$ by computing the Resource-Context-Similarity described in [5]. The latter is based on a frequency-based representation of tags in the vector space of all resources, in which similarity is computed by the cosine similarity. Because rarely used tags have very sparse vector representations, we restricted ourselves to the 10,000 most frequently used tags. Based on the resulting pairwise similarity values, we added an edge (t_1, t_2) to the edge list E_{sim} when the similarity was above a given threshold $min\_sim = 0.04$. This threshold was determined by inspecting the distribution of all similarity values. Table 1 summarizes the statistics of all tagging datasets.

Subsequently, we computed all term abstractness measures introduced in the previous chapter based on *DEL*, *COOC* and *SIM*, i.e. $\sqsupset^{DEL}_{freq}$, $\sqsupset^{DEL}_{entr}$, $\sqsupset^{COOC}_{dc}$, $\sqsupset^{COOC}_{bc}$, $\sqsupset^{COOC}_{cc}$, $\sqsupset^{SIM}_{bc}$, $\sqsupset^{SIM}_{cc}$ and $\sqsupset^{\mathbb{F}}_{subs}$.

5.3 Direct Evaluation Metric

As stated above, our grounding datasets contain information about concept subsumptions. If a concept c_1 subsumes concept c_2 (i.e. $(c_1, c_2) \in \geq_C$), we assume c_1 to be more abstract than c_2; as the taxonomic relation is transitive, we can infer $(c_1, c_2), (c_2, c_3) \in \geq_C \Rightarrow (c_1, c_3) \in \geq_C$ and hence that c_1 is also more abstract than c_3. In other words, thinking of the taxonomic relation as a directed graph, a given concept c is more abstract than all other concepts contained in the subgraph rooted at c. As we are interested in abstractness judgments about lexical items, we can consequently infer that concept labels for more abstract concepts are more abstract themselves. However, hereby we are facing the problem of polysemy: A given lexical item l can be used as a label for several concepts

[13] The data set is publicly available at `http://www.uni-koblenz-landau.de/koblenz/fb4/AGStaab/Research/DataSets/PINTSExperimentsDataSets/index_html`



(a) WordNet (b) YAGO

(c) DMOZ (d) WikiTaxonomy

Fig. 1: Grounding of each introduced term abstractness measure $\sqsupset^{\mathbb{S}}$ against four ground-truth taxonomies. Each bar corresponds to a term abstractness measure; the y-axis depicts the gamma correlation as defined in Equation 7. (Values for `cc_sim` and `bc_sim` in (d) are -0.05 and -0.005, resp.)

of different abstractness levels. Consequently, l has "several" abstractness levels, depending in which context it is used. As a most simple approach, which removes possible effects of word sense disambiguation techniques, we "resolve" ambiguity in the following way: The abstractness measure $\sqsupset^{\mathbb{O}} \subseteq L^C \times L^C$ on the vocabulary of a core ontology $\mathbb{O}$ is constructed according to

$$(l_1, l_2) \in \sqsupset^{\mathbb{O}} \Leftrightarrow (c_1, l_1) \in \mathcal{F} \wedge (c_2, l_2) \in \mathcal{F} \wedge (c_1, c_2) \in \geq_C \tag{6}$$

whereby $\mathcal{F}$ is the label assigment relation defined in Section 3 . Due to the polysemy effect described above, $\sqsupset^{\mathbb{O}}$ is not necessarily a partial order, as it may contain cycles. But despite this fact, $\sqsupset^{\mathbb{O}}$ contains the complete information which terms $l_i \in L^C$ are more abstract than other terms $l_j \in L^C$ according to the taxonomy of $\mathbb{O}$. Hence we can use it as a "ground truth" to judge the quality of a given term abstractness measure $\sqsupset^{\mathbb{S}}$.

We are interested how well $\sqsupset^{\mathbb{O}}$ correlates to $\sqsupset^{\mathbb{S}}$; picking up the idea of the *gamma rank correlation* [6], we define *concordant* and *discordant* pairs between $\sqsupset^{\mathbb{S}}$ and $\sqsupset^{\mathbb{S}}$ as follows: a pair of terms l and k is called concordant w.r.t. two partial orderings $\sqsupset, \sqsupset_*$, if they agree on it, i.e. $(l \sqsupset k \wedge l \sqsupset_* k) \vee (k \sqsupset l \wedge k \sqsupset_* l)$. It is called discordant if they disagree, i.e. $(l \sqsupset k \wedge k \sqsupset_* l) \vee (k \sqsupset l \wedge l \sqsupset_* k)$. Note that there may exist pairs which are neither concordant nor discordant.

Based on these notions, the gamma rank correlation is defined as

$$CR(\sqsupset, \sqsupset_*) = \frac{|C| - |D|}{|C| + |D|} \tag{7}$$

whereby C and D denote the set of concordant and discordant pairs, respectively.

In our case, $\sqsupset_*$ is not a partial ordering, but only a relation – which means that in the worst case, a pair l, k can be concordant and discordant at the same time. As is obvious from the definition of the gamma correlation (see Eq. 7), such inconsistencies lead to a lower correlation. Hence, our proposed method of "resolving" term ambiguity by constructing $\sqsupset^{\mathbb{O}}$ according to Eq. 6 leads to a lower bound of correlation. Figure 1 summarizes the correlation of each of our analyzed measures, grounded against each of our ground truth taxonomies. First of all, one can observe that the correlation values between the different grounding datasets differ significantly. This is most obvious for the DMOZ hierarchy, where almost all measures perform only slightly better than random guessing. A slight exception is the entropy-based abstractness measure $\sqsupset^{\mathbb{F}}_{entropy}$, which in general gives greater than 0.25 across all datasets. Another relatively constant impression is that the centrality measures based on the tag similarity graph (*cc_sim* and *bc_sim*) show a smaller correlation than the other measures. The globally best correlations are found for the WikiTaxonomy dataset, namely by the subsumption-model-based measure *subs*. Apart from that, the centrality measures based on the tag cooccurrence graph and the frequency-based measure show a similar behavior.

5.4 Derived measures

The grounding approach of the previous section gave a first impression of the ability of each measure to predict term abstractness judgments explicitly present in a given taxonomy. This methodology allowed only for an evaluation based on term pairs between which a connection exists in $\sqsupset^{\mathbb{O}}$, i.e. pairs where l_1 is either a predecessor or a successor of l_2 in the term subsumption hierarchy. However, our proposed measures make further distinctions among terms between which no connection exists within a taxonomy (e.g. the *freq* states that the most frequent term t is more abstract than *all* other terms). This phenomenon can probably also be found when asking humans – e.g. if one would ask which of the terms "art" or "theoretical computer science" is more abstract, most people will probably choose "art", even though both words are not connected by the is-a relation in (at least most) general-purpose taxonomies.

In order to extend our evaluation to these cases, we derived two straightforward measures from a taxonomy which allow for a comparison of the abstractness level between terms occurring in disconnected parts of the taxonomy graph. Because this approach goes beyond the explicitly encoded abstractness information, the question is justified to which extent it makes sense to compare the generality of completely unrelated terms, e.g. between "waterfall" and "chair". Besides our own intuition, we are not aware of any reliable method to determine

Table 2: Results from the user study.

Category	Number of classifications
One tag more general	41
Same level	11
Not comparable	154
Do not know one or two tags	3

when humans perceive the abstractness of two terms as *comparable* or not. For this reason, we validated the derived measures – namely (i) the shortest path to the taxonomy root and (ii)the number of subordinate terms – by an experiment with human subjects.

Shortest path to taxonomy root As stated above, most taxonomies are built in a top-down fashion, whereby more abstract terms are more likely to occur closer to the taxonomy root. Hence, a natural candidate for judging the abstractness of a term is to measure its distance to the root node. This corresponds to a ranking function $sp\_root(l)$, which ranks the terms l contained in a taxonomy in ascending order by the length of the shortest path between *root* and l.

Number of subordinate terms Another measure is inspired by Kammann et al. [16], who stated that "*the abstractness of a word or a concept is determined by the number of subordinate words it embraces[. . .]*". Given a taxonomy $\mathbb{O}$ and its comprised term subsumption relation $\sqsupset^{\mathbb{O}}$, we can easily determine the number of "sub-terms" by $subgraph\_size(l) = \text{card}\{(l, l') \in \sqsupset^{\mathbb{O}}\}$. We are aware that this measure is strongly influenced e.g. by fast-evolving domains like e.g. "mobile computing", whose rapid growth along with a strong expansion of the included vocabulary might lead to an overestimation of its abstractness level. This is another motivating reason for the user study presented in the next paragraph.

Validation by user study In order to check whether $sp\_root(l)$ and $subgraph\_size(l)$ correspond to human judgments of term abstractness, we performed an exemplary user study with 12 participants[14]. As a test set, we drew a random sample of 100 popular terms occurring in each of our datasets; for each term, we selected 3 candidate terms, taking into account cooccurrence information from the folksonomy *DEL*. The resulting 300 term pairs were shown to the each subject via a web interface[15], asking them to label the pair by one of 5 options (see Table 2)

We calculated Fleiss' κ [10] to get a closer look at the agreement of the study participants. In our experiment, $\kappa = 0.2836$ is indicating fair agreement. Table 2 shows the results of the number of classifications given that an agreement of 6 or more participants signalizes significant agreement. The relatively high number

[14] students and staff from two IT departments

[15] `http://www.kde.cs.uni-kassel.de/benz/generality_game.html`

Table 3: Accuracy of the taxonomy-derived abstractness measures.

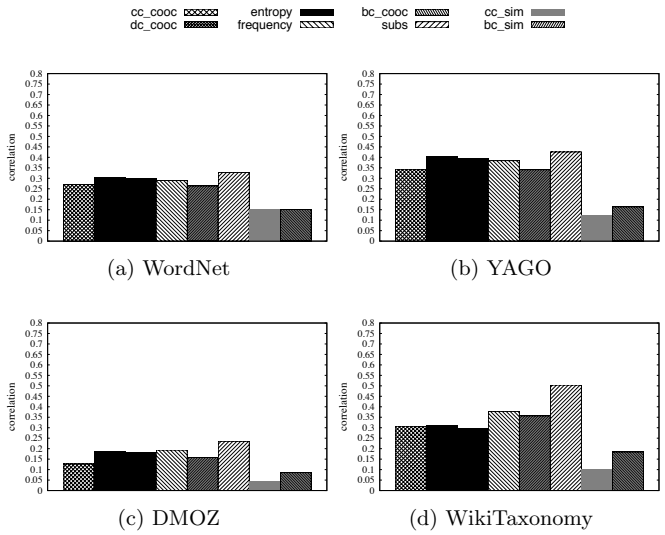
	Wordnet	Yago	DMOZ	WikiTaxonomy
sp_root	0.94	0.42	0.88	0.45
subgraph_size	0.94	0.96	0.8	0.87

of "not comparable" judgments show that even with our elaborate filtering, the task of differentiating abstractness levels is quite difficult. Despite this fact, our user study provided us with a well-agreed set of 41 term pairs, for which we got reliable abstractness judgments. Denoting these pairs as $\sqsupset_{manual}$, we can now check the accuracy of the term abstractness measures introduced by *sp_root* and *subgraph_size*, i.e. the percentage of correctly predicted pairs. Table 3 contains the resulting accuracy values. From our sample data, it seems that the subgraph size (i.e. the number of subordinate terms) is a more reliable predictor of human abstractness judgments. Hence, we will use it for a more detailed grounding of our folksonomy-based abstractness measures.

The ranking function *subgraph_size* naturally induces a partial order $\sqsupset^{\mathbb{O}}_{subgraph\_size}$ among the set of lexical items present in a core ontology $\mathbb{O}$. In order to check how close each of our introduced term abstractness measures correlate, we computed the *gamma correlation coefficient* [6] between the two partial orders (see Eq. 7). Figure 2 shows the resulting correlations. Again, the correlation level between the datasets differs, with DMOZ having the lowest values. This is consistent with the first evaluation based solely on the taxonomic relations (see Figure 1). Another consistent observation is that the measure based on the tag similarity network (*bc_sim* and *cc_sim*) show the weakest performance. The globally best value is found for the subsumption model, compared to the WikiTaxonomy (0.5); for the remaining conditions, almost all correlation values lie in the range between 0.25 and 0.4, and correlate hence weakly.

5.5 Discussion

Our primary goal during the evaluation was to check if folksonomy-based term abstractness measures are able to make reliable judgments about the relative abstractness level of terms. A first consistent observation is that measures based on frequency, entropy or centrality in the tag cooccurrence graph do exhibit a correlation to the abstractness information encoded in gold-standard-taxonomies. One exception is DMOZ, for which almost all measures exhibit only very weak correlation values. We attribute this to the fact that the semantics of the DMOZ topic hierarchy is much less precise compared to the other grounding datasets; as an example, the category `Top/Computers/Multimedia/Music_and_Audio/Software/Java` does hardly imply that `Software` "is a kind of" `Music_and_Audio`. WordNet on the contrary subsumes the term *Java* (among others) under taxonomically much more precise parents: `[...] > communication > language > artifical language > programming language > java` The same holds for Yago, and the WikiTaxonomy was also built with a strong focus on *is-a* relations [21]. This is actually an interesting observation: Despite the fact that both DMOZ and Delicious were built for similar



(a) WordNet (b) YAGO

(c) DMOZ (d) WikiTaxonomy

Fig. 2: Grounding of the term abstractness measure $\sqsupset^{\mathbb{S}}$ against $\sqsupset^{\mathbb{O}}_{subgraph\_size}$ derived from four ground-truth taxonomies. Each bar corresponds to a term abstractness measure; the y-axis depicts the gamma correlation as defined in Equation 7.

purposes (namely organizing WWW references), the implicit semantics within Delicious resembles more closely to well-established semantic repositories than to the bookmark-folder-inspired hierarchical organization scheme of DMOZ.

Another consistent observation is that abstractness measures based on tag similarity graphs (as used e.g. by [13]) perform worst through all experimental conditions. This is consistent with observations in our own prior work [5], where we showed that distributional similarity measures (like the one used in this paper or by [13]) induce connections preferably among tags having the same generality level. On the contrary, applying e.g. centrality measures to the "plain" tag cooccurrence graph yield better results. Hence, a justifiable conclusion is that tag-tag cooccurrence encodes a considerable amount of "taxonomic" information.

But this information is not solely present in the cooccurrence graph – also a probabilistic model of subsumption [22] yields good results in some conditions, especially when grounding against the taxonomy-derived *subgraph_size* ranking. We attribute this to the fact that both measures (the subsumption model and the subgraph size) are based on the same principle, namely that a term is more general the more other terms it subsumes.

Apart from that, even the simplest approach of measuring term abstractness by the mere frequency (i.e. the number of times a tag has been used) already exhibits a considerable correlation to our gold-standard taxonomies. This has an

interesting application to the *popularity/generality problem*: Our results point in the direction that popular tags are on average more abstract (or more general) than less frequently used ones. In summary, the interpretation of our results can be condensed in two statements: First, folksonomy-based measures of term abstractness do exhibit a partially strong correlation to well-defined semantic repositories; and second, the abstractness level of a given tag can be approximated well by simple measures.

6 Conclusions

In this paper, we performed a systematic analysis of folksonomy-based term abstractness measures. To this end, we first provided a common terminology to subsume the notion of term abstractness in folksonomies and core ontologies. We then contributed a methodology to compare the abstractness information contained in each of our analyzed measures to established taxonomies, namely WordNet, Yago, DMOZ and the WikiTaxonomy. Our results suggest that centrality and entropy measures can differentiate well between abstract and concrete terms. In addition, we have provided evidence that the tag cooccurence graph is a more valuable input to centrality measures compared to tag similarity graphs in order to measure abstractness. Apart from that, we also shed light on the *tag generality vs. popularity* problem by showing that in fact, popularity seems to be a fairly good indicator of the "true" generality of a given tag. These insights are useful for all kinds of applications who benefit from a deeper understanding of tag semantics. As an example, tag recommendation engines could take generality information into account in order to improve their predictions, or folksonomy navigation facilities could offer a new direction of browsing towards more general or more specific directions. Finally, our results inform the design of algorithms geared towards making the implicit semantics in folksonomies explicit.

As next steps, we plan to apply our measures to identify generalists and specialists in social tagging systems. A possible hypothesis hereby is that specialists use a more specific vocabulary whereas generalists rely mainly on abstract tags.

Acknowledgments. We would like to thank Dr. Denis Helic and Beate Krause for fruitful discussions during the creation of this work. 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 WebZubi project funded by BMBF and the VENUS project funded by Land Hessen.

References

1. Allen, R., Wu, Y.: Generality of texts. In: Digital Libraries: People, Knowledge, and Technology. Lecture Notes in Computer Science, Springer, Heidelberg (2010)
2. Benz, D., Hotho, A., Stumme, G.: Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In: Proc. of WebSci2010. Raleigh, NC, USA (2010)

3. Boszak, E. et al: KAON – Towards a Large Scale Semantic Web. In: Proc. EC-Web 2002, LNCS, vol. 2445, pp. 304–313. Springer (2002)
4. Brandes, U., Pich, C.: Centrality estimation in large networks. I. J. Bifurcation and Chaos 17(7), 2303–2318 (2007)
5. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic analysis of tag similarity measures in collaborative tagging systems. In: Proc. of the 3rd Workshop on Ontology Learning and Population (OLP3). pp. 39–43. Patras, Greece (July 2008)
6. Cheng, W., Rademaker, M., Baets, B.D., Hüllermeier, E.: Predicting partial orders: Ranking with abstention. In: ECML/PKDD. LNCS, vol. 6321. Springer (2010)
7. Cimiano, P., Hotho, A., Staab, S.: Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research (JAIR) 24, 305–339 (2005)
8. Clark, J., Paivio, A.: Extensions of the Paivio, Yuille, and Madigan (1968) norms. Behavior Research Methods, Instruments, & Computers 36(3), 371 (2004)
9. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
10. Fleiss, J., et al.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
11. Golder, S., Huberman, B.A.: Usage patterns of collaborative tagging systems. Journal of Information Science 32(2), 198–208 (April 2006)
12. Henschel, A., Woon, W.L., Wächter, T., Madnick, S.: Comparison of generality based algorithm variants for automatic taxonomy generation. In: Proc. of IIT09. pp. 206–210. IEEE Press, Piscataway, NJ, USA (2009)
13. Heymann, P., Garcia-Molina, H.: Collaborative creation of communal hierarchical taxonomies in social tagging systems. Tech. Rep. 2006-10, CS dep. (April 2006)
14. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: SIGIR '08: Proc. of the 31st annual Int'l ACM SIGIR conference on Research and development in information retrieval. pp. 531–538. ACM (2008)
15. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Information retrieval in folksonomies: Search and ranking. In: The Semantic Web: Research and Applications. LNAI, vol. 4011, pp. 411–426. Springer (2006)
16. Kammann, R., Streeter, L.: Two meanings of word abstractness. Journal of Verbal Learning and Verbal Behavior 10(3), 303 – 306 (1971)
17. Körner, C., Benz, D., Hotho, A., Strohmaier, M., Stumme, G.: Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In: Proc. of WWW2010. pp. 521–530. ACM (2010)
18. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishing, Boston (2002)
19. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In: International Semantic Web Conference. pp. 522–536. LNCS, Springer (2005)
20. Paivio, A., Yuille, J.C., Madigan, S.A.: Concreteness, imagery, and meaningfulness values for 925 nouns. Journal of Experimental Psychology 76 (1968)
21. Ponzetto, S.P., Strube, M.: Deriving a large-scale taxonomy from wikipedia. In: AAAI. pp. 1440–1445. AAAI Press (2007)
22. Schmitz, P.: Inducing ontology from flickr tags (2006)
23. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: 16th international World Wide Web conference (WWW 2007) (2007)
24. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge Univ Pr (1994)
25. Zhang, Q.: Fuzziness - vagueness - generality - ambiguity. Journal of Pragmatics 29(1), 13 – 31 (1998)

4. Conclusions

Past research analyzed the motivation behind tagging by using empirical studies. The publications contained in this thesis establish a quantitative evaluation to enable the distinction of two different types of tagging motivation - categorizers and describers - and present the associated work. For differentiation purposes, measures that inspect a user's personomy are introduced. Furthermore, experiments were conducted that investigate the effects user groups with different motivations have on the resulting folksonomies. The findings of these experiments show that tagging motivation has influences on knowledge extraction tasks such as social classification and the inference of semantics from folksonomies.

4.1. Results and contributions

To conclude and summarize this work, the answers to the research questions defined in Section 1.3 are given:

RQ1 - What kinds of tagging motivation can be identified in social tagging systems?

In addition to tagging motivation already presented by other research this work identifies two different kinds of user motivation - categorizers and describers. Users driven by description tend to characterize their resources in a detailed and elaborate manner while introducing new tags frequently and therefore not maintaining a stable tag vocabulary. Categorizers on the other hand establish a personal tag vocabulary to which they stick, often re-use tags and try to keep semantic ambiguities between them to a minimum.

RQ2 - Is it possible to measure tagging motivation automatically?

To differentiate the two types of user motivation this thesis presents a number of different measures that analyze the personomy of a given user. These measures range from ratios between a user's tag assignments and resources to the analysis of the user's encoding process of documents via tags. The advantages of the measures are *content agnosticism*, *language independence* and the usage of *individual user characteristics* as input data for the computation instead of the complete folksonomy. The presented measures indicate that the approximation of tagging motivation is feasible by inspecting simple statistical properties of a user's tagging vocabulary. This way expensive and time consuming empirical studies can be circumvented.

RQ3 - How does tagging motivation vary within and across systems?

To answer this research question the previously introduced measures were applied on snapshots of different social tagging systems such as Delicious, Flickr, BibSonomy, Diigo and others to study how tagging motivation differs inside of single and across multiple tagging systems. For comparison purposes two artificial reference datasets that characterize extreme categorization and description behavior were generated. The results of this analysis showed that there is indeed a variety in tagging motivation in and across the examined systems. Although users exhibiting extreme behavior could be observed, in reality the tagging motivation is a mixture of both types - categorization and description. However, it was found that systems such as MovieLens tend to have users more driven towards categorization whereas Delicious has more describers in their user base.

RQ4 - What effects does tagging motivation have on knowledge extraction tasks like social classification and semantic extraction?

For both types of tagging motivation - categorizers as well as describers - tasks were identified which constitute better results than the other group. Groups of users exhibiting describing motivation are better suited for automated inferring of semantic structures in folksonomic systems. Other research indicates that groups of users driven by categorizing provide good data for classifying books into the category schemes of the Library of Congress Classification and the Dewey Decimal Classification scheme. These results show that user behavior has an influence on the resulting structures within a folksonomy - information that is useful for researchers and system designers of social tagging systems.

By answering the four research questions the contributions of this dissertation can be outlined.

4.1.1. Contributions

The contributions of the presented thesis are the following:

- The *differentiation between two types of tagging motivation* - describers and categorizers (Section 3.4).
- *Introduction of quantitative measures* to distinguish between categorizers and describers and the subsequent evaluation (Sections 3.4 and 3.6).
- *Analysis of how different motivations influence the resulting folksonomy* and the effects on tag recommendation (Section 3.5).
- *Experiments that evaluate how categorizing and describing behavior affects knowledge extraction tasks* such as social classification and the emergence of semantics in social tagging systems (Sections 3.7 and 3.8).

4.2. Implications of this work

Based on our findings, developers of social tagging systems are able to measure and differentiate tagging motivation into two different types - categorizers and describers.

A direct implication is the possible implementation of tagging interfaces that adapt automatically based on the behavioral characteristics of a user and assist in a personalized way. We have already shown that the distinction of the two proposed user groups can improve the performance of tag recommendation, but the integration of this feature into an existing tag recommendation mechanism is still an open question.

Furthermore, during the context of this work we found evidence for a casual link between the pragmatics and the semantics of a social tagging system. Informed by these results, designers and developers of such systems can influence and guide users towards a specific tagging behavior by implementing user interfaces and tag recommendation mechanisms that support a type chosen by system architects.

4.3. Limitations and future work

Although this thesis presents a way of automatically inferring user motivation, the quantitative analysis of user behavior to get insight into tagging motivation is still at an early stage. Some research directions that can be envisioned from the current status of this work are sketched in the following.

Impact of existing mechanisms on tagging behavior
While we observed evidence that existing mechanisms such as tag recommendation interfaces and auto completion have impact on the behavior of a user, there is still a lot to learn about how these supporting features influence users' decisions when applying tags. Results of such work might be especially useful for system designers focused on driving users towards a particular kind of tagging behavior.

What impact does the type of annotated resources have on tagging behavior?
We showed that tagging motivation varies within and across different types of social tagging systems. However, this thesis does not give answers on how the type of a resource (e.g. pictures, videos or scientific articles) makes an impact on tagging motivation and subsequent user behavior in folksonomic systems.

Tagging behavior and navigability in social tagging systems
Another interesting research aspect is how user behavior influences navigability in social tagging systems. We already explored navigational aspects in previous research (cf. [Trattner et al., 2011] and [Helic et al., 2012]), but did not include user behavior as a perspective for our experiments. An interesting future experiment is the analysis of which user groups support efficient navigation in social systems. A possible hypothesis is that users driven by categorization establish more efficient paths for browsing within these platforms since they use a consistent vocabulary.

Nomenclature

cf.	confer
DDC	Dewey Decimal Classification
etc.	et cetera
HTML	HyperText Markup Language
LCC	Library of Congress Classification System
URL	Uniform Resource Locator
vs	versus

List of Figures

Bibliography

[Benz et al., 2010] Benz, D., Hotho, A., and Stumme, G. (2010). Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference (WebSci10)*, Raleigh, NC, USA.

[Benz et al., 2011] Benz, D., Körner, C., Hotho, A., Stumme, G., and Strohmaier, M. (2011). One tag to bind them all: Measuring term abstractness in social metadata. In Antoniou, G., Grobelnik, M., Simperl, E., Parsia, B., Plexousakis, D., Pan, J., and Leenheer, P. D., editors, *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*, Heraklion, Crete.

[Cattuto et al., 2007] Cattuto, C., Loreto, V., and Pietronero, L. (2007). Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464.

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

[Coates, 2005] Coates, T. (2005). Two cultures of fauxonomies collide... http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxonomies_collide/, last accessed Janurary 26th 2012.

[Dellschaft and Staab, 2008] Dellschaft, K. and Staab, S. (2008). An epistemic dynamic model for tagging systems. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 71–80, New York, NY, USA. ACM.

[Dewey, 1876] Dewey, M. (1876). *A Classification and subject index for cataloguing and arranging the books and pamphlets of a library.* Amherst, Massachusetts.

[Farooq et al., 2007] Farooq, U., Kannampallil, T. G., Song, Y., Ganoe, C. H., Carroll, J. M., and Giles, L. (2007). Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*, pages 351–360, New York, NY, USA. ACM.

[Golder and Huberman, 2006] Golder, S. A. and Huberman, B. A. (2006). The structure of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208.

[Grahsl, 2010] Grahsl, H.-P. (2010). Pragmatic analysis of tagging motivation in social tagging systems. Master's thesis, Graz University of Technology.

[Halpin et al., 2007] Halpin, H., Robu, V., and Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA. ACM.

[Hammond et al., 2005] Hammond, T., Hannay, T., Lund, B., and Joanna, S. (23.05.2005). Social bookmarking tools (i): A general review.

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

[Heckner et al., 2008] Heckner, M., Neubauer, T., and Wolff, C. (2008). Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types. In *Proceedings of the 2008 ACM workshop on Search in social media*, SSM '08, pages 3–10, New York, NY, USA. ACM.

[Helic et al., 2012] Helic, D., Körner, C., Granitzer, M., Strohmaier, M.,

and Trattner, C. (2012). Navigational efficiency of broad vs. narrow folksonomies. In *Proceedings of the 23nd ACM conference on Hypertext and hypermedia*.

[Helic and Strohmaier, 2011] Helic, D. and Strohmaier, M. (2011). Building directories for social tagging systems. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 525–534, New York, NY, USA. ACM.

[Helic et al., 2011] Helic, D., Strohmaier, M., Trattner, C., Muhr, M., and Lerman, K. (2011). Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 417–426, New York, NY, USA. ACM.

[Helic et al., 2010] Helic, D., Trattner, C., Strohmaier, M., and Andrews, K. (2010). On the navigability of social tagging systems. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 161–168, Washington, DC, USA. IEEE Computer Society.

[Hong et al., 2008] Hong, L., Chi, E. H., Budiu, R., Pirolli, P., and Nelson, L. (2008). Spartag.us: a low cost tagging system for foraging of web content. In *Proceedings of the working conference on Advanced visual interfaces*, AVI '08, pages 65–72, New York, NY, USA. ACM.

[Hotho et al., 2006] Hotho, A., Jäschke, R., Schmitz, C., and Stumme, G. (2006). Bibsonomy: A social bookmark and publication sharing system. In de Moor, A., Polovina, S., and Delugach, H., editors, *Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, Aalborg, Denmark. Aalborg University Press.

[Kern et al., 2010] Kern, R., Körner, C., and Strohmaier, M. (2010). Exploring the influence of tagging motivation on tagging behavior. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'10, pages 461–465, Berlin, Heidelberg. Springer-Verlag.

[Kipp and Campbell, 2006] Kipp, M. E. I. and Campbell, D. G. (2006). Patterns and inconsistencies in collaborative tagging systems: An ex-

amination of tagging practices. In *Annual General Meeting of the American Society for Information Science and Technology, Austin, Texas (US), 3-8 November 2006*.

[Körner et al., 2010a] Körner, C., Benz, D., Strohmaier, M., Hotho, A., and Stumme, G. (2010a). Stop thinking, start tagging - tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th International World Wide Web Conference (WWW 2010)*, Raleigh, NC, USA. ACM.

[Körner et al., 2010b] Körner, C., Kern, R., Grahsl, H. P., and Strohmaier, M. (2010b). Of categorizers and describers: An evaluation of quantitative measures for tagging motivation. In *21st ACM SIGWEB Conference on Hypertext and Hypermedia (HT 2010)*, Toronto, Canada. ACM.

[Körner and Strohmaier, 2010] Körner, C. and Strohmaier, M. (2010). A call for social tagging datasets. *SIGWEB Newsl.*, pages 2:1–2:6.

[Ley and Seitlinger, 2010] Ley, T. and Seitlinger, P. (2010). A cognitive perspective on emergent semantics in collaborative tagging: The basic level effect. In Cena, E. F., Dattolo, A., Kleanthous, S., Tasso, C., Vallejo, D. B., and Vassileva:, J., editors, *CEUR Workshop Proceedings of the International Workshop on Adaptation in Social and Semantic Web (SASWeb2010)*, volume 590, pages 13–18.

[Marlow et al., 2006] Marlow, C., Naaman, M., Boyd, D., and Davis, M. (2006). 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. ACM.

[Mathes, 2004] Mathes, A. (2004). Folksonomies - cooperative classification and communication through shared metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, last accessed February 2nd 2012.

[Mika, 2007] Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15.

[Nov et al., 2009] Nov, O., Naaman, M., and Ye, C. (2009). Motivational

, structural and tenure factors that impact online community photo sharing. *MIS Quarterly*, pages 138–145.

[Plangprasopchok and Lerman, 2008] Plangprasopchok, A. and Lerman, K. (2008). Constructing folksonomies from user-specified relations on flickr. *CoRR*, abs/0805.3747.

[Quintarelli, 2005] Quintarelli, E. (2005). Folksonomies: power to the people. http://www.iskoi.org/doc/folksonomies.htm, last accessed January 29th 2012.

[Rader and Wash, 2008] Rader, E. and Wash, R. (2008). Influences on tag choices in del.icio.us. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 239–248, New York, NY, USA. ACM.

[Sen et al., 2006] Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M., and Riedl, J. (2006). tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, CSCW '06, pages 181–190, New York, NY, USA. ACM.

[Shen and Wu, 2005] Shen, K. and Wu, L. (2005). Folksonomy as a complex network. cite arxiv:cs/0509072.

[Sinha, 2005] Sinha, R. (2005). A cognitive analysis of tagging. http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/ - last accessed February 28th 2012.

[Smith, 2008] Smith, G. (2008). *Tagging: People-powered Metadata for the Social Web (Voices That Matter)*. New Riders Press, 1 edition.

[Stirling, 2007] Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719.

[Strohmaier et al., 2010] Strohmaier, M., Körner, C., and Kern, R. (2010). Why do users tag? detecting users' motivation for tagging in social tagging systems. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

[Trattner et al., 2011] Trattner, C., Körner, C., and Helic, D. (2011). Enhancing the navigability of social tagging systems with tag taxonomies. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, pages 18:1–18:8, New York, NY, USA. ACM.

[Wal, 2005] Wal, T. V. (2005). Explaining and showing broad and narrow folksonomies. Blog post. http://www.personalinfocloud.com/2005/02/explaining_and_.html - last accessed February 1st 2012.

[Wal, 2007] Wal, T. V. (2007). Folksonomy coinage and definition. http://vanderwal.net/folksonomy.html last accessed Jan 15th 2012.

[Wash and Rader, 2007] Wash, R. and Rader, E. (2007). Public bookmarks and private benefits: An analysis of incentives in social computing. *Proceedings of the American Society for Information Science and Technology 2007*.

[Xu et al., 2006] Xu, Z., Fu, Y., Mao, J., and Su, D. (2006). Towards the semantic web: Collaborative tag suggestions. In *Proceedings of the Collaborative Web Tagging Workshop at the WWW 2006*, Edinburgh, Scotland.

[Zubiaga et al., 2011] Zubiaga, A., Körner, C., and Strohmaier, M. (2011). Tags vs shelves: from social tagging to social classification. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 93–102, New York, NY, USA. ACM.

A. Complete list of own publications

A.1. Journal articles

- Markus Strohmaier, Denis Helic, Dominik Benz, Christian Körner, Roman Kern (2011) *Evaluation of Folksonomy Induction Algorithms*, 22. In Transactions on Intelligent Systems and Technology ACM TIST V (N).
- Mark Kröll, Christian Körner, Markus Strohmaier (2010) *ITAG: Automatically Annotating Textual Resources with Human Intentions*, 333-342. In JETWI Vol 2, No 4 (2010): Special Issue: Recommender Systems for Web Intelligence 2 (4).

A.2. Conference proceedings

- Karin Schöfegger, Christian Körner, Philipp Singer, Michael Granitzer (2012) *Learning User Characteristics From Social Tagging Behavior.* In Conference on Hypertext and Hypermedia (Short Paper).
- Denis Helic, Christian Körner, Michael Granitzer, Markus Strohmaier, Christoph Trattner (2012) *Navigational Efficiency of broad vs narrow folksonomies.* In Conference on Hypertext and Hypermedia.
- Peter Kraker, Christian Körner, Kris Jack, Michael Granitzer (2012) *Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization.* In 1st International Workshop on Large Scale Network Analysis, WWW 2012.
- Claudia Wagner, Silvia Mitter, Christian Körner, Markus Strohmaier (2012) *When social bots attack: Modeling susceptibility of*

users in online social networks, 8. In Proceedings of the 2nd Workshop on Making Sense of Microposts (MSM'2012).

- Christoph Trattner, Christian Körner, Denis Helic (2011) *Enhancing the Navigability of Social Tagging Systems with Tag Taxonomies*. In 11th International Conference on Knowledge Management and Knowledge Technologies.

- Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, Markus Strohmaier (2011) *One Tag to Bind Them All: Measuring Term Abstractness in Social Metadata*. In Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011).

- Arkaitz Zubiaga, Christian Körner, Markus Strohmaier (2011) *Tags vs Shelves: From Social Tagging to Social Classification*. In HT '11: Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia.

- Roman Kern, Christian Körner, Markus Strohmaier (2010) *Exploring the Influence of Tagging Motivation on Tagging Behavior*. In European Conference on Research and Advanced Technology for Digital Libraries (ECDL).

- Christian Körner, Roman Kern, Hans-Peter Grahsl, Markus Strohmaier (2010) *Of categorizers and describers: an evaluation of quantitative measures for tagging motivation*, 157-166. In Conference on Hypertext and Hypermedia.

- Christian Körner, Dominik Benz, Andreas Hotho, Markus Strohmaier, Gerd Stumme (2010) *Stop thinking, start tagging: tag semantics emerge from collaborative verbosity*, 521-530. In International World Wide Web Conference.

- Markus Strohmaier, Christian Körner, Roman Kern (2010) *Why do users tag? Detecting users' motivation for tagging in social tagging systems*. In ICWSM 2010 Conference on Weblogs and Social Media.

- Markus Strohmaier, Mark Kröll, Christian Körner (2009) *Automatically annotating textual resources with human intentions*, 355-356. In Conference on Hypertext and Hypermedia.

- Markus Strohmaier, Mark Kröll, Christian Körner (2009) *Intentional query suggestion: making user goals more explicit during search*, 68–74. In Proceedings of the 2009 workshop on Web Search Click Data.

A.3. Miscellaneous/Additional publications

- Christian Körner, Markus Strohmaier (2010) *A call for social tagging datasets*. ACM SIGWEB Newsletter, Winter Issue, January 2010
- Christian Körner (2009) *Understanding the Motivation behind Tagging*, First Place at ACM Student Research Competition, Conference of Hypertext and Hypermedia 2009