

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

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

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

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

Copyright 2010 ACM 978-1-4503-0041-4/10/06 ...\$10.00.

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 person (containing all tags and resources of a user) 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 dis-

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

inction introduced above is theoretic, we would expect that 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.

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

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 or-

3d 9/11 Berlusconi **IT** Web2.0 advertising agency alternative
amarcord animation anthropology architecture art asia
astronomy berlusconi **blog** brushes climate cms comics
compatibility **css** culture design docs doomsday
economics energy environment experimental
flash flashdev free fun geniality **graphics** hacks
health history humor icons identity illustration
india inspiration interaction inutilities iran iraq italy
javascript job logos mac mafia mainstream media
misteriditalia movies music navigation nerd news pattern
photography php pisos pixel **politics** portfolio
print privacy recipes religion rights satellite science seo
shockwave shop **society** stock streetart tcpa template thc torrent
travel tutorial tv type typography **utilities** video war
web2.0 webdesign webdev women world wtf
zeitgeist

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.

phaned tags frequently and 0 identifies users who maintain a more consistent vocabulary. Considering the categorizer - describer paradigm this would mean that categorizers would 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.

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

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:

!read !video Books Didaktik GUI Hotels acce accessibility admin aggregation agile ai air ajax amazon
analyze ant apache api apple **apps** art audio austria auto aws backup barcamp barcode bayes
behaviour berlin bildungsungleichheiten blogs **book** bookmarklets books brand
browser business cache **cakephp** cakephp calendar canvas capistrano charts classes cms
cocoa collaboration conference continuousintegration contracts cooking copyright cruisecontrol crypt **css** cursor
datasource debug **del.icio.us** deployment design dev devhouse domain dos download
dr dsl ebook ec2 eclipse economy editor elearning election email experiment facebook financial finanzen firebug
firefox firmware flash flex fly fonds forms framework freelancer fritz fun gallery game gateway gears gettext
git google **googlemaps** greasemonkey **gtd** gui handy highlighting hiagi hosting htaccess **html** hulu i18n
iPod ia icons ide **ie** info information inhaltsstoffe interieur interview invoice **iphone** ipod ischgl iso jQuery **jobs**
jquery **js** jslint jsp juristisches keyboard lastfm latex learn lebensmittel legal library lifehacks logo lokal
mac **magazine** mail map maps marketing mathematics media message migration mobile wandern
movie msql music mvc **mysql** münchen nd news openid os p2p pattern patterns paypal performance
phone photo **php** plugin pm png podcast politics post print private process programming
prototyping proxy psychologie qm **qs** ratings read **readlist** recipes reference remember rente republica
research ressources restaurant rhetorics **rezepte** rss ruby ruhrgbiet rücken s3 safari sandra scalability
schach school schwammverhalten screenshot scrum search seo series shop **shopping** ski skype slides soa
social **software** sound spam sql startup stats sterben study subtitles subversion svn tax **test**
testing textmate texts thinkpad time tool **tools** trac travel tutorial tv typography ubuntu unobtrusive
unterricht urheberrecht urlaub usability utf8 vacation viamento.info video visualization voip wandern
weather **web** **web2.0** webcam **webdesign** westwing wetter widgets wiki windows wohnung word
wordpress work xdebug xhtml xp zend zitae

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.

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

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:

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

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 by the following formula:

$$overlap = 1 - \frac{|R_u|}{|TAS_u|} \quad (5)$$

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 Snowball¹ stemmer. For normalization purposes we relate the resulting absolute intersection size to the cardinality of the set of title words.

$$ttr = \frac{|T_u \cap TW_u|}{|TW_u|} \quad (6)$$

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.

¹<http://snowball.tartarus.org/>

²Details 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.

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.

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

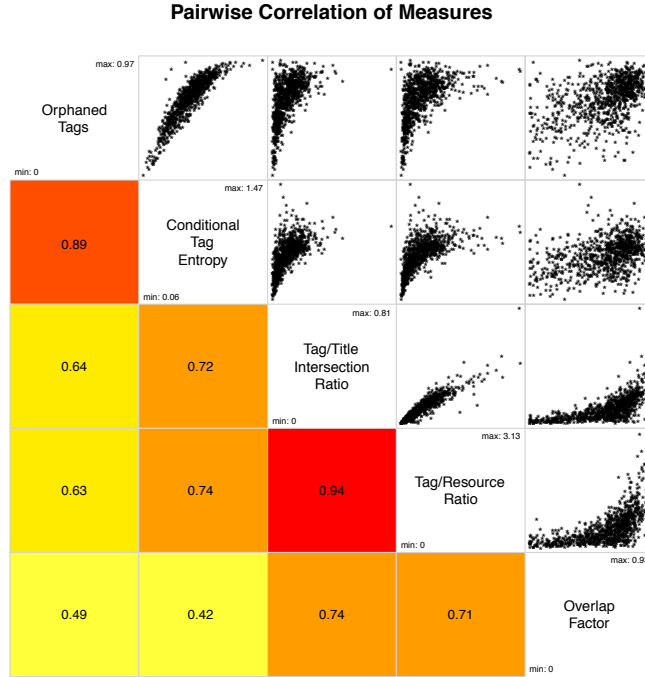


Figure 3: Spearman rank correlation of the measures for the Del.icio.us dataset (correlation values in the lower left and pairwise distribution in the upper left). All measured were developed based on different intuitions and capture different aspects of the describers and categorizers behavior, still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

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 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 Co-

	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

hen’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

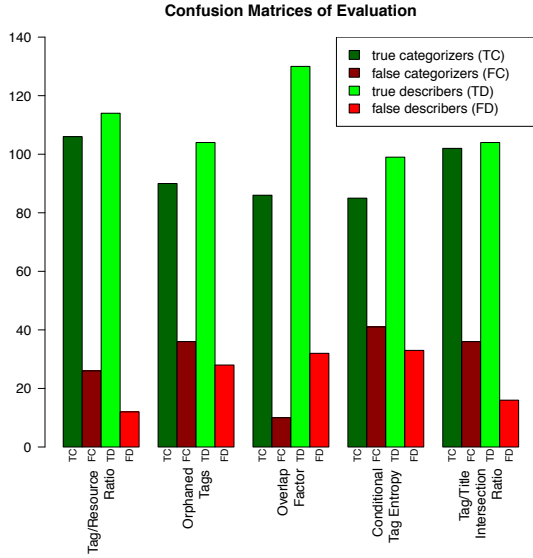


Figure 4: Confusion matrices of the evaluation

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 decide 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 Delicio.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 Descriptor, FC...False Categorizer and FD...False Descriptor. 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 in-

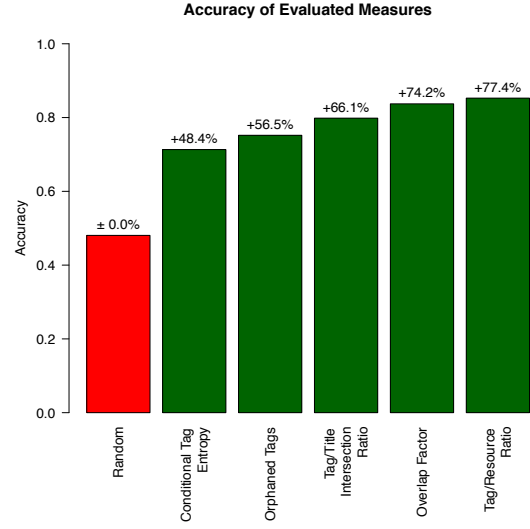


Figure 5: Accuracy for the different measures resulting from the user study

fluence 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.

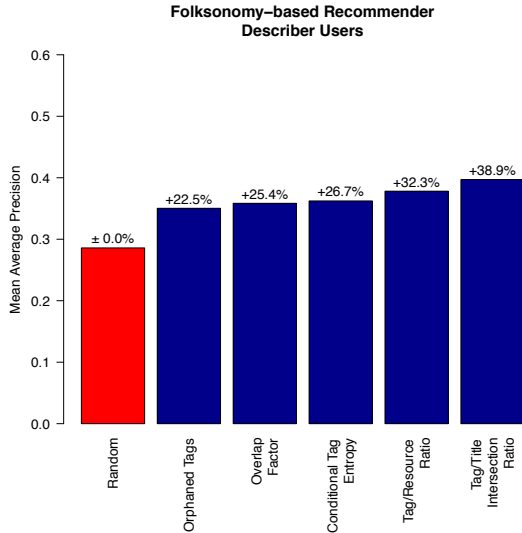


Figure 6: MAP for the set of describers influenced by the folksonomy-based recommender. All differently defined groups of describers are influenced by the folksonomy-based recommender.

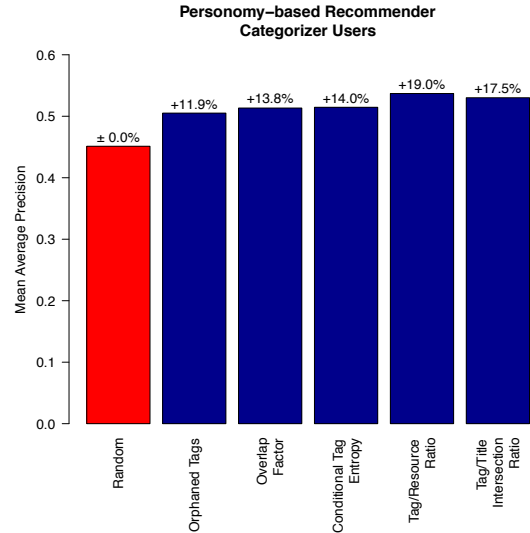


Figure 7: MAP for the set of categorizers influenced by the personomy-based recommender. All differently defined groups of categorizers are influenced by the personomy-based recommender.

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 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,

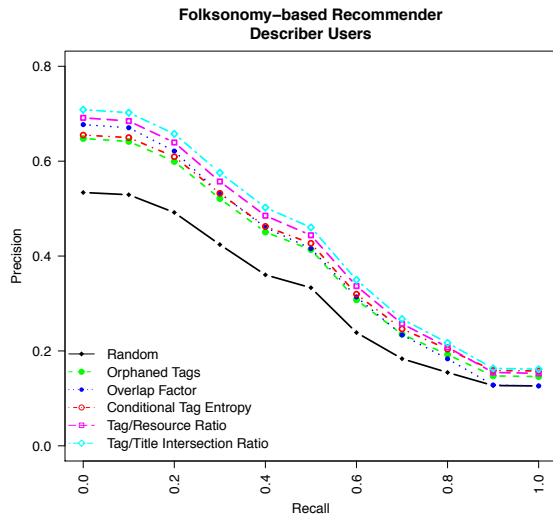


Figure 8: Precision/Recall curve for the describer users in combination with the folksonomy-based recommender. Across different recall levels, the folksonomy-based recommender influences all differently defined describer groups.

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 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

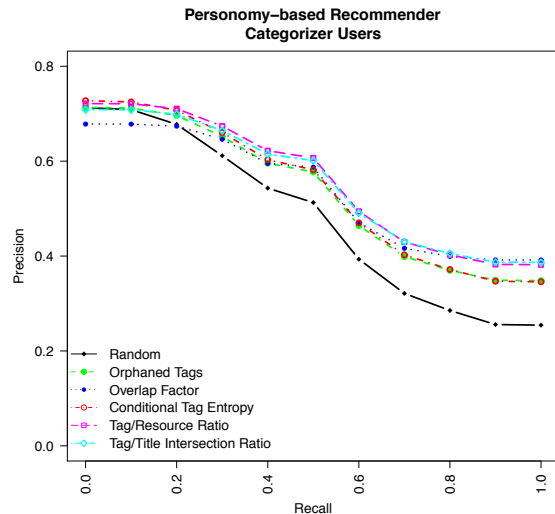


Figure 9: Precision/Recall for the categorizer/personomy-based recommender configuration. Across different recall levels, the personomy-based recommender influences all differently defined categorizer groups.

our quantitative evaluation, using recommender algorithms to simulate latent influence, show that the motivation behind tagging has a significant effect on tagging behavior. The 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 fauxnomies collide. http://www.plasticbag.org/archives/2005/06/two_cultures_of_fauxnomies_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.