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.

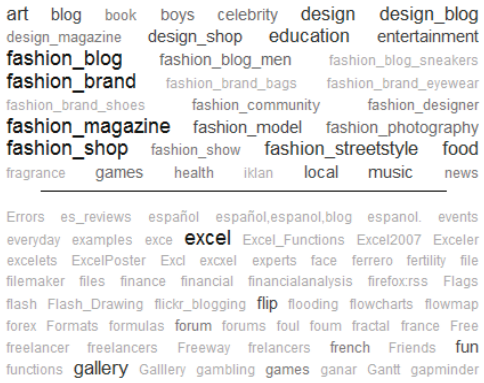

Figure 1: Examples of tag clouds produced by users who are driven by different motivations for tagging: categorization (top) vs. description (bottom)

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/

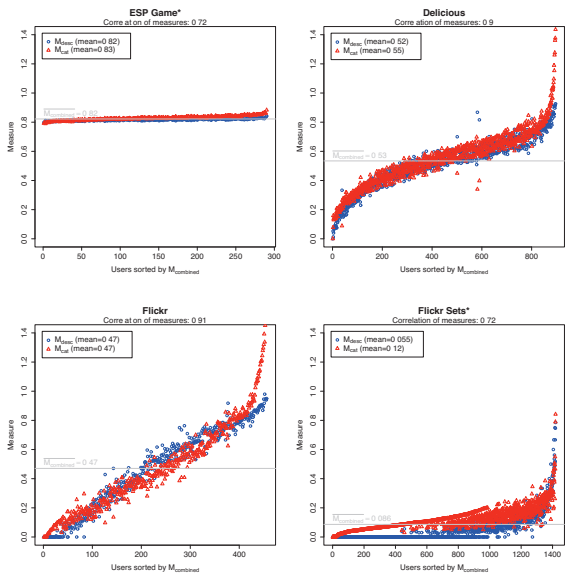

Figure 2: M_{desc} and M_{cat} at $|R_u| = 500$ for 4 datasets, including Pearson correlation and the mean value for $M_{combined}$.

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

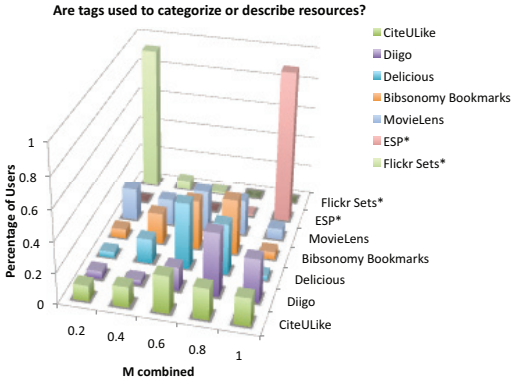

Figure 3: $M_{combined}$ at $|R_u| = 500$ for 7 different datasets, binned in the interval [0.0 .. 1.0]. The two back rows reflect opposite extreme behaviors.

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. (to appear).

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.