The impact of resource title on tags in collaborative tagging systems

Marek Lipczak
Faculty of Computer Science
Dalhousie University
Halifax, Canada, B3H 1W5
lipczak@cs.dal.ca

Evangelos Milios
Faculty of Computer Science
Dalhousie University
Halifax, Canada, B3H 1W5
eem@cs.dal.ca

ABSTRACT

Collaborative tagging systems are popular tools for organization, sharing and retrieval of web resources. Their success is due to their freedom and simplicity of use. To post a resource, the user should only define a set of tags that would position the resource in the system's data structure – folksonomy. This data structure can serve as a rich source of information about relations between tags and concepts they represent. To make use of information collaboratively added to folksonomies, we need to understand how users make tagging decisions. Three factors that are believed to influence user tagging decisions are: the tags used by other users, the organization of user's personal repository and the knowledge model shared between users. In our work we examine the role of another potential factor – resource title. Despite all the advantages of tags, tagging is a tedious process. To minimize the effort, users are likely to tag with keywords that are easily available. We show that resource title, as a source of useful tags, is easy to access and comprehend. Given a choice of two tags with the same meaning, users are likely to be influenced by their presence in the title. However, a factor that seems to have stronger impact on users' tagging decisions is maintaining the consistency of the personal profile of tags. The results of our study reveal a new, less idealistic picture of collaborative tagging systems, in which the collaborative aspect seems to be less important than personal gains and convenience.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Experimentation

Keywords

collaborative tagging, folksonomies, modelling

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

1. INTRODUCTION

Collaborative tagging systems allow users to create public repositories of web resources. Each resource is entered into the system in the form of a *post* which consists of the resource, the user posting it and a short description of the resource given by the user. The key to the success of collaborative tagging systems lies in the complete lack of description formalism. To describe a resource the user enters a set of free-form keywords, called *tags*. Tagging turns a cumbersome classification problem, in which each resource should be assigned a place in a hierarchy of classes, into an unstructured categorization problem, in which each resource is related to a set of loose ad-hoc user-defined categories [10, 11]. Despite the fact that users have complete freedom in choosing their tags, it is widely believed that constant interaction with posts of other users leads to collaborative actions and emergence of a pseudo taxonomy, called *folksonomy*. Although, the term folksonomy refers to the pseudo hierarchy of tags, it is often used to refer to the complete data structure created in collaborative tagging systems, or even as a synonym of collaborative tagging systems. A folksonomy is formally represented as a tripartite graph of hyper-edges connecting resources, users and tags [15, 20]. In this paper the representation is simplified to a projection of the graph to a set of tags connected to a given resource (or a user). The set, called *profile*, contains all tags that were ever used in a post of a resource (*resource profile*) or by user (*user profile*). together with the number of occurrences of each tag (*frequency*).

1.1 Tagging models

Tagging is a complex process which involves actions of a large community of users. To make it easier to understand we usually view this process as a combination of tagging models. The three most frequently discussed tagging models are the collaborative, personal and shared knowledge model. The *collaborative model* [5, 7, 10, 23] assumes that, while tagging, users take into consideration tags attached to the same resource by other users. This can happen directly when a user adopts a resource from someone else or indirectly when a user assigns tags suggested by a tag recommendation system, which usually draws the recommended tags from the resource profile. This model is the basis of the folksonomy self-organization assumption. The *personal model* [19, 22, 23] assumes that a user treats the collaborative tagging system as a personal repository of web resources, ignoring its collective character. In this case, the main aim of the user is to re-use personal tags to organize an individual library of

resources. The *shared knowledge model* [7, 10, 11] assumes that all users comprehend the content of the tagged resource in a similar way, hence they should come up with a similar set of tags to describe it as they are pulling the tags from a shared repository of descriptions that capture the semantics of the resource. The collaborative and personal models are in obvious contradiction and are quite easy to characterize. On the other hand, the role of shared knowledge model is hard to identify because of the vague nature of the resource semantics and the fact that its effect can be confused with that of the two other models.

There is no doubt that the behaviour described by all presented models is present in real user actions. However, together with the theoretical models, we should consider some practical aspects of tagging process. A key issue not addressed by the collaborative and personal models is that users would only spend the effort to maintain high quality of tags, in a collaborative or personal sense, if they see obvious advantages. Despite the benefits of tagging for organizational and information retrieval purposes, tagging is a burden, which implies that users are likely to spend the least time and effort on it. The obvious way to ease the tagging process is using tags proposed by tag recommendation systems. Tag recommenders are available in many collaborative tagging systems, yet measuring their impact on the tagging process is hard, as each system uses its own recommendation algorithm and interface, which in addition undergo frequent updates. In our work we decided to focus on another, more explicit, element that can play an important role in tagging process – the resource title. We were motivated by recent research on tag recommendation, which revealed relatively high overlap between tags and words extracted from the resource title, making the title an important part of tag recommendation system [8]. The title, as a dense resource description, is likely to contain useful keywords. In addition, the title is usually visible during tagging. As a result, the title is a convenient source of tags that may ease the burden of tagging. Using words from the resource title as tags may be considered as a realization of the shared knowledge model. The set of terms that describe a resource well is limited and the same terms could be chosen independently by the author of the title and the user assigning tags. However, this is true only to some extent for the following reasons. First, resource authors often trade title clarity and precision for attractiveness. Second, pulling words out of the context of a well-formed sentence can lead to lower tag quality, and be a reason for the high tag disorder observed in folksonomies.

1.2 Study outline

The objective of our work was to examine how strong is the relation between resource title and tags used to describe this resource. We wanted to *(a)* find evidence that the choice of tags is influenced by the title; *(b)* relate the tagging decisions which are not in line with the title to the basic tagging models and *(c)* identify mechanisms that prevent users from using title words as tags. During our study we examined the following hypotheses:

HYPOTHESIS 1. *Occurrences of terms as tags and as title words are related.*

We found that for the majority of tags the occurrences of terms as tags and title words cannot be claimed to be independent.

HYPOTHESIS 2. *Occurrence in the title is a key factor of the popularity of the tag in resource profile.*

Comparison of title words and the most frequent tags from the resource profile showed large overlap between title and popular tags. This observation lessens the impact of collaborative behaviour of users on tagging. Although the results show that title is potentially one of the key factors of term popularity it is clear that it cannot be the only factor that impacts the tagging process.

HYPOTHESIS 3. *Given two tags that convey the same meaning, users are more likely to pick the tag that can be found in the title.*

To test this hypothesis we used a short and precise list of pairs of terms that as tags can be considered synonymous. The pairs are singular and plural form of a noun (e.g., *blog* and *blogs*), which convey, as tags, the same meaning. The observation of resources that represent similar concepts gave us evidence that the choice between synonymous tags is indeed affected by their occurrence in the title. However, the importance of the title should not be overestimated. The occurrence of the word in the title does not lead to a domination of the resource profile by a single tag form. Resources are constantly tagged with both forms of the tag. In fact, the occurrence of a term form in the title does not even determine that the form will be used more frequently than the other. The popularity for each individual pair of tags is biased towards one of the forms (e.g., *blog*) and the occurrence of the other form of the term (*blogs*) in the title is not always able to change this pattern. To find factors that decrease the impact of the title on tagging decisions we checked if the popularity bias and constant use of both tag forms can be explained by any of the three tagging models.

HYPOTHESIS 3.1. *The collaborative tagging model can explain the tag form popularity bias and constant use of both forms of the tag.*

Close observation of tagging patterns for individual resources confirmed the weak impact of the collective character of tagging, making the collaborative tagging model unlikely to explain these characteristics.

HYPOTHESIS 3.2. *The shared knowledge model can explain the tag form popularity bias and constant use of both forms of the tag.*

We found that the popularity of one form of a term as a tag is reflected in its popularity as a title word. It suggests that the popularity bias is likely to be due to the common knowledge model shared between the resource authors and users assigning the tags.

HYPOTHESIS 3.3. *The personal model can explain the tag form popularity bias and constant use of both forms of the tag.*

We found evidence that the influence of personal model is a strong factor that prevents one of the tag forms to dominate the resource profile. In addition, we found that users are less likely to be affected by the resource title while picking the tag form when the tag is playing an important role in their profile.

2. RELATED WORK

The first models of tagging behaviour to explain observed folksonomy characteristics (differences in popularity of tags, stabilization of tag proportions and power-law in resource profiles) were based on generative processes which assumed a common vocabulary of tags from which users draw their decisions [10, 5, 11]. They all assumed collaborative behaviour of users. Recently, Dellschaft and Staab [7] extended the collaboration based generative model considering the impact of shared knowledge vocabulary to match additional folksonomy characteristics (e.g., sub-linear growth of tags). In contrast, Rader and Wash [22] showed that user's tagging decisions are more affected by the need of personal profile organization than the impact of collaborative suggestions. These results were confirmed by the work of Wetzker et al. [26], who suggested that users develop their personal vocabulary and proposed a method to map it to the general folksonomy vocabulary. Evidence that statistical characteristics of folksonomies (e.g., emerging power-law tag distribution in resource profiles) are not caused by collaborative behaviour of users was provided by Bollen and Halpin [3]. Based on the results of a user study they concluded that power-law distributions emerge independently of the availability of collaborative suggestions. Krause et al. [18] showed that folksonomies and so called *logsonomies*, which are data structures created based on search log data, have similar characteristics. The similarity occurs despite the fact that good tags are not likely to be good query terms and vice-versa [12]. It may suggest that patterns observed for folksonomies can be in fact typical for any kind of tripartite data structure of users, resources and keywords, even if there is no collaboration between users. This leads to the conclusion that a more general model (e.g., shared knowledge model) could be an explanation of folksonomy characteristics. All these studies focused on some characteristics of folksonomies and tried to find an explanation for them. In our work we decided to approach the problem from a different angle. Starting with an element that is likely to have impact on tagging decisions (resource title), we tried to describe its influence on the folksonomy characteristics and test the strength of models that are believed to shape them.

The interest in demonstrating the importance of resource title for collaborative tagging systems has been rather moderate in the literature. Most of the work is related to the tag recommendation task [17, 19, 24], in which resource title was recognized as a rich and precise source of tag recommendations. Similar conclusions were drawn by Heymann et al. [12] who measured the overlap of information represented by tags and website content (including title) to determine the usefulness of tags for web search. Figueiredo et al. [9] examined the quality of title (and other post fields) for other information retrieval and data mining tasks. The title turned out to be the most descriptive feature; however, it does not generalize the information about the resource as well as tags, which are superior for classification tasks.

3. DATASETS

In our experiments we used datasets from three collaborative tagging systems: *Delicious*[1] – a repository of website bookmarks, *CiteULike*[2] – a repository of references to scientific publications and *BibSonomy*[3] which combines the functionality of both [14]. As the *BibSonomy* dataset is still fairly small we decided not to report exact results for this dataset, because of their low accuracy. In general, the results obtained for *BibSonomy* data confirm the findings for *Delicious* and *CiteULike* dataset. Because of space limitations we decided to focus on these two datasets only.

3.1 Delicious dataset

Despite the fact that *Delicious* does not make its dataset publicly available for research purposes, its size and popularity makes it a frequent object of interest. We had access to two Delicious snapshots which we refer to (according to their origin) as *Delicious MPII*[4] and *Delicious TUB DAI-Labor*[5]. *Delicious MPII* snapshot [2] was crawled in July 2007 using a snowball sampling method [1] starting with one of the largest (non-spammer) user profiles and following "fan" links available in *Delicious*. The snapshot contains profiles of over 13,000 users (see Table 1). *Delicious TUB DAI-Labor* [25] was crawled in April 2008, using tag based snowball sampling to create a list of active users. The snapshot contains over 900,000 profiles of the most active users found in the first round of crawling. Both snapshots contain the basic information about a post – user, resource and tags. In addition, *Delicious MPII* dataset contains the title of posted website and *Delicious TUB DAI-Labor* contains the posting date. Both title and time-stamp were needed to run our experiments, which means we had to combine the datasets to create an intersection that we refer to as *Delicious* dataset. Although matching was not trivial as user ID in both datasets was obfuscated, it was feasible thanks to their resulting from a similar approach to crawling, highly overlapping time span of posts and large size of user profiles (Fig. 1(c)). To combine the datasets we matched tag-based and resource-based profiles of users from both datasets, which in most cases gave strong one-to-one overlap. The matching process revealed that part of the posts in *Delicious MPII* snapshot missed some of the tags. In such cases we decided to use all the tags that for the given post could be found in *Delicious TUB DAI-Labor* dataset. As a result *Delicious* dataset used in the study contained more unique tags than *Delicious MPII* dataset (Table 1).

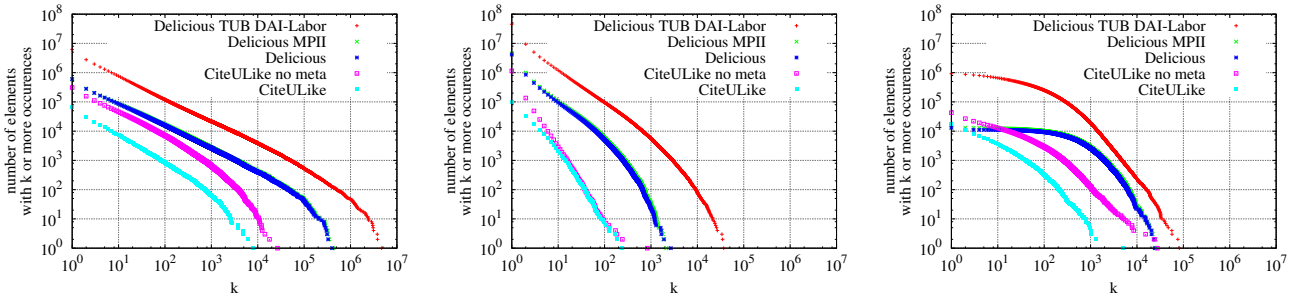
3.2 CiteULike dataset

CiteULike makes its dataset available for research purposes daily. The dataset used in this study was downloaded on December 15, 2009. Unfortunately, CiteULike does not provide resource information, including the resource title. To make the dataset applicable to our experiments, we manually downloaded the titles of the most frequently posted resources using the *CiteGeist*[6] feature. Using the search interface and combination of queries addressing tags and title field we manually downloaded metadata for resources that were tagged with or contained in their title one of the 102 tags used in the experiments on *CiteULike* dataset (see Section. 4.2 for more information about tags). During this process no information about real user IDs or profiles was revealed to us. In general, for all datasets we managed to

[1]http://delicious.com/
[2]http://www.citeulike.org/
[3]http://www.bibsonomy.org/
[4]Max-Planck-Institut für Informatik – http://www.mpi-inf.mpg.de/
[5]TU Berlin - DAI Laboratory – http://www.dai-labor.de/
[6]http://www.citeulike.org/citegeist

	posts	tags			resources		users	
	total	total	unique	top freq	unique	top freq	unique	top freq
Delicious TUB DAI-Labor	123,248,945	410,700,267	6,201,251	4,680,275	45,333,406	36,998	922,652	81,992
Delicious MPII	9,847,813	31,739,921	588,635	421,596	4,469,945	2,139	13,240	25,755
Delicious	8,890,876	29,807,506	601,547	399,927	4,172,960	2,673	13,079	24,176
CiteULike no meta	1,420,922	4,941,571	311,588	26,735	1,148,163	836	42,876	27,553
CiteULike	200,291	619,958	65,087	8,050	96,965	242	17,693	4,923

Table 1: Statistics of *Delicious* and *CiteULike* datasets used in the study, together with base datasets (*top freq* is the number of occurrences of a most frequent element).



(a) Tags (b) Resources (c) Users

Figure 1: Cumulative frequency distributions for *Delicious* and *CiteULike* datasets (with base datasets). *Delicious* and *Delicious MPII* distributions are nearly identical and overlapping. We observe the expected behaviour of tags distribution [4] – the power-law distribution with cut-off for high frequency tags. The resource distribution reveal the main drawback of *CiteULike* dataset and folksonomies in general – low number of resources posted frequently. The effect of crawling can be noticed in user distribution for both base *Delicious* datasets, where we only have the information about the tail of the distribution (users with large number of posts).

preserve the anonymity of users and avoided the need of additional crawling.

3.3 Preliminaries

Additional features of collaborative tagging systems and freedom of use contribute to the noisy character of folksonomies. If not handled carefully, the noise is likely to strongly bias the results of experiments. Before running the study we considered the following sources of noise, trying to understand and if needed reduce their impact on the results of our study.

Spammers – although *Delicious TUB DAI-Labor* dataset is known to be strongly biased by spammers [25], the combination of the two *Delicious* datasets used in our experiments is free from this problem thanks to "fan" based crawling process used to obtain *Delicious MPII* dataset. We examined CiteULike data manually and have not found symptoms of spamming behaviour [21].

Imported posts – collaborative tagging systems allow users to import their resources from external repositories (e.g., browser bookmarks) or other collaborative tagging systems. Such posts can strongly bias user profiles and obfuscate tagging patterns because they introduce large number of automatically generated tags. We eliminated posts which contained tags likely associated with imported posts (e.g., "firefoxbookmarks", "bibtex_import"). The users may choose not to use such tags. To overcome that problem we tried to expose and remove imports by finding large groups of resources posted by a single user in a short time period. This technique is not effective for *Delicious* data, because while importing posts from browser's bookmarks folder, *Delicious* copies the time-stamp of creating the bookmark and subfolder names as tags, making these posts hard to distinguish from real posts.

Tag recommendation – The *Delicious* tag recommendation interface, which was in use at the time that the posts used in our study were formulated, contained a short list of recommended tags (likely containing the most frequent tags from the resource profile) and a long list of user profile tags [22]. The title was not a part of the recommendation list although it was present in the posting window as the description of a resource. *CiteULike* added tag recommendation feature recently. It offers only tag suggestions extracted from the resource content (title and abstract if available). The recommender builds multi-word phrases and recommends them as single tags. Independently, the full title is visible when tags are entered into the system and can influence the choice of tags. As this property is frequent in collaborative tagging systems, we consider it more as their feature rather than bias.

4. EXPERIMENTS

To examine the potential influence of the title on the choice of tags we ran a series of experiments. The experiments can be divided into two sets. The objective of the first set was to confirm the relation between the title words and tags. In the second set we focused on word pairs, which as tags convey the same meaning. The objective was to ob-

serve if preference for one of the words was affected by its occurrence in the title.

4.1 Overlap of title and resource profile

HYPOTHESIS 1. *Occurrences of terms as tags and as title words are related.*

To check if the occurrence of a term as a tag is related to its occurrence as a title word, we examined terms that were used at least 100 times as a tag or could be found in a resource title of at least 100 posts (36,558 terms for *Delicious* dataset and 2,155 terms for *CiteULike* dataset). This threshold was chosen to remove the potential noise caused by low-frequency terms. For each term we checked in how many posts the term can be found (a) as a tag, but not in the title, (b) in the title but not as a tag and (c) both as a tag and in the title. We extracted terms for which the number of posts in each of the three sets was at least five (17,821 terms for *Delicious* dataset and 1,532 terms for *CiteULike* dataset). We ran the *Pearson's chi-square test of independence* for each of these terms. In each case the *null* hypothesis (independence of tags and title words) was rejected with high confidence $p < 0.0001$. Hence, for these terms we are able to confirm that use of a term as a tag is related to its occurrence in the title.

We manually browsed the list of terms that were rejected from the experiment because of an insufficient number of samples. We focused on the terms, which, despite being popular as tags, could not be found in the title. We found that a significant part of these tags (30% for *Delicious* dataset and 54% for *CiteULike* dataset) matched `(w+W)+w` regular expression pattern, where `w` stands for a letter and `W` stands for a non-letter character used to separate words. These tags are complex terms composed of two or more words (e.g., "social_networks"). In this case it is likely that the relation between the title and tags exists as well, but is too complex to be captured by our experiment.

To get quantitative information about the overlap of title words and tags, we processed all posts in both datasets counting the number of times a tag can be found in the title of tagged resource. The experiment shows that 15% of tags in *Delicious* dataset and 26% of tags in *CiteULike* dataset can be found in the title. The large difference between the datasets is likely to be caused by the different tag recommender and character of the resources. The title of a web page is usually shorter and less descriptive than the title of a scientific publication, hence the former is likely to provide fewer terms that can be useful as tags.

HYPOTHESIS 2. *Occurrence in the title is a key factor of the popularity of the tag in resource profile.*

The potential importance of the title in the formulation of resource profile was revealed by the second experiment in which we took the profiles of frequently posted resources and calculated the likelihood of a title word being highly ranked in the profile (number of times the tag with rank k was found as the title word, divided by the number of tested resources). We set the threshold of accepting the resource as frequently tagged at 100 for *Delicious* dataset and 20 for *CiteULike* dataset. The choice of the threshold value followed the work by Heymann et al. [13], who showed that the list of the top 100 tags in the resource profile originates mainly in the first 100 posts. Unfortunately because of a low number of frequently posted resources we had to lower the threshold for *CiteULike* dataset. To reduce the bias caused by the variance in the number of posts per resource, we decided to use only the first 100 (or 20) posts to build the resource profile. For 40% of the tested resources for the *Delicious* dataset (50% for the *CiteULike* dataset) the top ranked tag in the profile was found in the title (Fig. 2). The probability of having a tag-title co-occurence rapidly decreases with the rank of the tag in the profile, which shows that title contains few high quality words that are used as tags frequently. On the other hand, the cumulative ratio of title words being used as top k tags is constantly growing with the increasing value of k, even for high k. Possibly these words are not good descriptors of the resource and they were used as tags only because they were noticed in the title. On average 60% of title words can be found among the top 100 or 40 tags of resource profiles for *Delicious* and *CiteULike* dataset respectively (Fig. 2).

The results of this experiment shed new light on the process of formulation of resource profile and its collaborative character which is generally assumed to be the main factor that determines the popularity of tags. The importance of the influence of other user decisions on the tagging process has already been questioned [3, 22], but here we present a potential alternative to the social influence. Instead of being influenced by others, users could be directly influenced by the content of the resource, specifically its title.

4.2 Synonymous tags

To observe the impact of resource title on tags, we used a set of pairs of terms, which can be used completely interchangeably to tag a resource. For simplicity we refer to them as synonymous tags, however, two synonymous tags do not have to be synonyms in natural language, which is the case in our study. Given a pair of synonymous tags we could observe the context of using them as tags to determine the sources of information or procedures that impact the choice between them.

The pair of synonymous tags, that we decided to focus our attention on, is a singular and plural form of the same noun. The fact, that these two forms used as tags convey the same meaning, was pointed out in previous work [6, 26], here we discuss the problem in more details. Most of the tags used in folksonomies are nouns, which is natural given that the aim of the tagging process is categorization of resources [11]. To categorize the resource, the noun can be used in singular or plural form to indicate that the resource (e.g., *blog*) belongs to a given category (e.g., *blogs*). We examined the list of the one thousand most frequent tags to find the popular (singular, plural) pairs of tags. It resulted in 96 pairs for the *Delicious* dataset and 51 pairs for the *CiteULike* dataset. These pairs were used in all the following experiments. Because of space limitations we present only the list of top ten pairs, sorted by the frequency of the more frequent form, for each dataset (Table 2). To confirm that two forms of the same term convey the same meaning when used as a tag, we looked at the resources for which one of the two forms was used at least 10 times. We then compared the sets of resources associated with the two forms of the same tag. The *Jaccard similarity coefficient*[7] [16] av-

[7]Jaccard similarity coefficient is defined for two sets as the size of their intersection divided by the size of their union

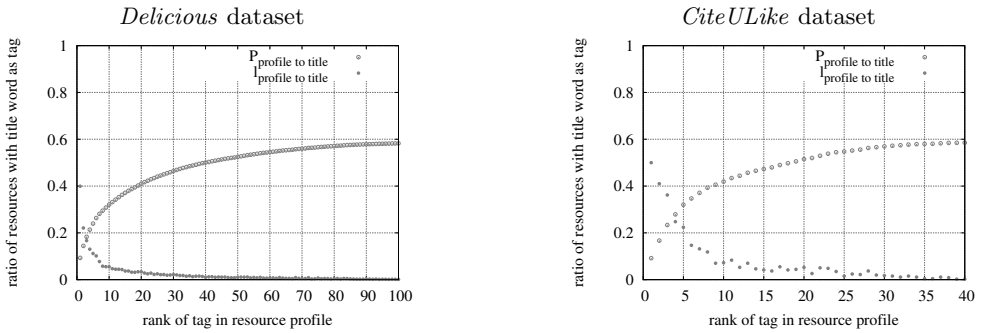


Figure 2: The overlap of title and resource profile: $l_{profile\ to\ title}$ – what percentage of profile tags with rank k can be found in the title, $P_{profile\ to\ title}$ – what percentage of top k profile tags can be found among title words.

eraged over all pairs is 0.83 for the *Delicious* dataset and 0.76 for the *CiteULike* dataset. The large number of (singular, plural) pairs among the most frequent tags and the high overlap between the resources described by the two forms of the same term agree with the intuition that such pairs can be viewed as functional synonyms. Focusing on the pairs of singular and plural forms of the same noun has an additional advantage in our study. Although they can be used interchangeably as tags it is not the case in natural language. Often the form of a term is determined by the longer phrase in which it is used (see Table 3 for examples). The situation in which a concept can be represented by both forms of a term, but the form that is used as a tag follows the form found in the title would be clear evidence that tags are influenced by the resource title. We used this idea in the following experiments, observing resource and user profiles in which one or two forms of a synonymous tag pair could be found.

4.2.1 The impact of resource title on resource profile

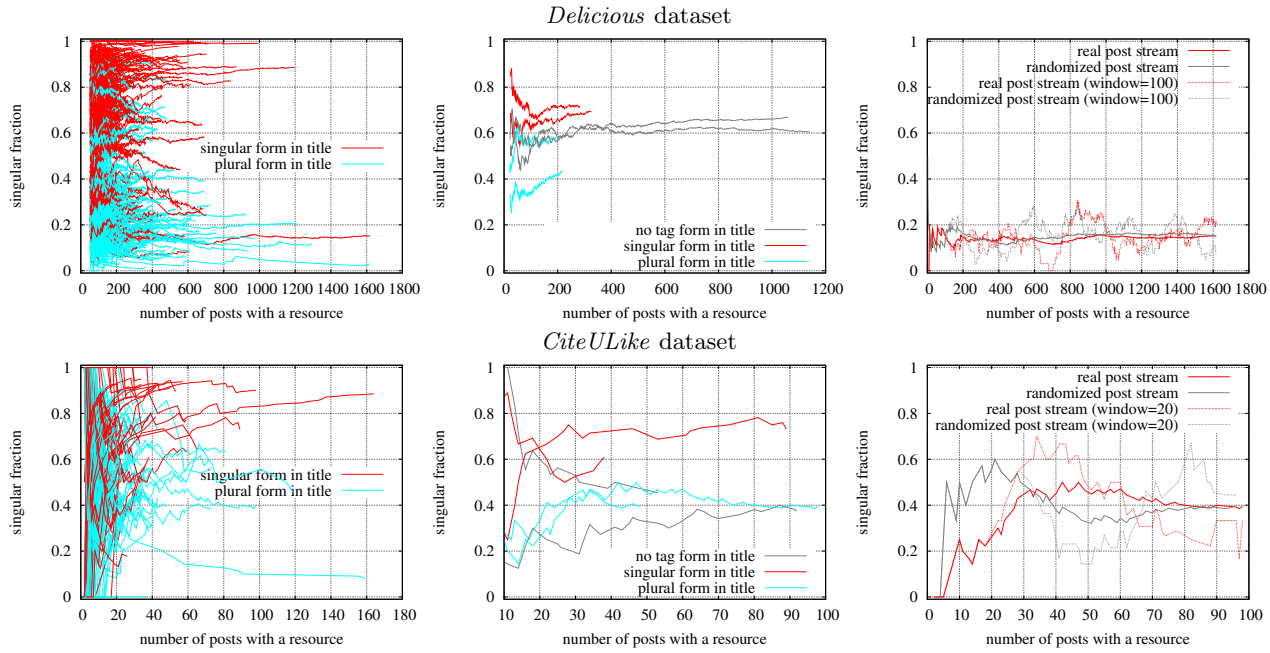
HYPOTHESIS 3. *Given two tags that convey the same meaning, users are more likely to pick the tag that can be found in the title.*

Knowing that words from the title are likely to be frequently used in the profile of a resource (Section 4.1), we decided to trace post streams of resources, to observe how the use of selected tags changes in time. A post stream [7] is a sequence of posts ordered by time-stamps. In our experiment we limited the stream to posts with a specific resource. We selected resources for which one of the two forms of (singular, plural) pair was frequently used as a tag (threshold of 20 or 5 uses for *Delicious* and *CiteULike* dataset respectively). Each frequently tagged resource for each tested pair was traced separately. Whenever one of the two tag forms was used, the fraction of singular tags among both singular and plural tags (*singular fraction* or sf) was recorded. If the occurrence of the word in the title has a direct impact on the choice of a tag we should observe it in the value of the *singular fraction*. The presence of the singular form of the tag should make the fraction high, whereas the presence of plural form should make it low.

The confirmation of the hypothesis can be found in the visualization of traces of the resources (Fig. 3(a)). We adapted the visualization method used in [10] and [5]. The *singular fraction* for a cumulated profile of each resource is presented as a single trace as a function of time, measured by the number of posts associated with the resource. The colour coding shows that the form of the title word is correlated with the dominant form of the tag in most cases. For most tested pairs (93% for *Delicious* dataset and 94% for *CiteULike* dataset) the average *singular fraction* calculated for full profiles of resources with the singular form of the term in the title is higher than the average fraction calculated for resources with plural form in the title. The form of the term as title word boosts its frequency as a tag. To confirm that the title has the determinant role in impacting the choice of

	singular frequency	(rank)	plural frequency	(rank)
Delicious dataset				
blog(s)	303973	(**4**)	146377	(19)
tool(s)	54152	(84)	266422	(**8**)
art(s)	237611	(**9**)	4497	(781)
video(s)	206044	(**11**)	17295	(258)
tutorial(s)	128669	(**24**)	52251	(88)
tip(s)	4052	(853)	106433	(**36**)
book(s)	43546	(106)	103169	(**38**)
game(s)	38411	(120)	102812	(**39**)
article(s)	80000	(**55**)	36120	(127)
wiki(s)	69394	(**66**)	5069	(694)
CiteULike dataset				
review(s)	27579	(**1**)	998	(910)
human(s)	13215	(10)	22420	(**2**)
animal(s)	3333	(173)	15952	(**4**)
model(s)	13563	(**7**)	11716	(16)
protein(s)	13300	(**9**)	7079	(50)
network(s)	12737	(**11**)	10105	(21)
method(s)	4828	(99)	9343	(**28**)
gene(s)	8402	(**35**)	3033	(198)
genetic(s)	6575	(53)	7629	(**44**)
cell(s)	7322	(**48**)	4381	(115)

Table 2: Top ten pairs out of the list of synonymous tags used in the study. Frequency and rank are calculated based on tags distribution (Fig. 1(a)). The terms are sorted by the frequency of the more frequent form. Pairs of singular and plural forms of tag are frequent in folksonomies. There is no general rule for more popular form of a tag.



(a) post streams of all resources traced in the experiment (resources with no tag form in title omitted for clarity)

(b) resources traced for tag *blog* (*Delicious*) and tag *network* (*CiteULike*) – two most frequent of each kind

(c) post stream of an example resource and its randomized version, with sliding window results

Figure 3: Results for resource profile tracing.

the tag form we would expect that among tags of a synonumous pair the majority of tags has the same form as the word in the title ($sf_{plural} < 0.5 < sf_{singular}$). Such clear division between *singular fraction* value for the resources with singular/plural form of the same tag in the title was observed for a small fraction of pairs only (25%, e.g., *download(s)*, for *Delicious* dataset and 22%, e.g., *network(s)* for *CiteULike* dataset). For these pairs the dominant form of the tag in the resource profile depends on the form that can be found in the resource title, even if both forms convey the same meaning (Fig. 3(b) bottom, and Table 3). The other pairs are strongly biased towards one of the forms (e.g., *blog* is the dominant form in *blog(s)* pair for *Delicious* dataset), and, even though the occurrence of the less frequent form in the title influences the choice of a tag, it is often used in a minority of posts for a given resource (Fig. 3(b) top, and Table 3). Although the title is a factor that impacts the choice of the tag form, in most cases, its impact is not strong enough to overcome the popularity bias caused by some other factors.

It is interesting to notice that the value of *singular fraction* rarely reaches boundary values. Both forms of the tag are constantly added to the system and the ratio between them seems to stabilize over time. This behaviour is analogous to the results of experiments by Golder and Huberman [10], where the stabilization was claimed to be a result of two factors: imitation of other user tags (collaborative model) and shared knowledge. We investigated these two potential explanations. To complete the picture, we investigated the impact of the title on user profiles to check if the observed characteristics can be explained by the personal model.

resource title	sf
Delicious dataset	
Blog Software Breakdown	0.68
(...) Create your **Blog** Now – FREE	0.68
Blog software comparison chart	0.67
(...) Where **Blogs** Meet Maps	0.58
CiteULike dataset	
Folksonomy as a complex **network**	0.73
Exploring complex **networks**	0.39
Complex **networks**: Structure and dynamics	0.39
Statistical mechanics of complex **networks**	0.38

Table 3: Example of resources related to the concept of blogging (*Delicious*) and complex networks (*CiteULike*). The form of term found in the title boosts its frequency as a tag.

HYPOTHESIS 3.1. *The collaborative tagging model can explain the tag form popularity bias and constant use of both forms of the tag.*

To look closer at the stabilization process, we redesigned the experiment and calculated the *singular fraction* in a sliding window of posts. We set the size of the window to 100 for the *Delicious* dataset, as Golder and Huberman [10] suggested this is the number of posts after which the stabilization is observed. The window size had to be reduced to 20 for the *CiteULike* dataset because of insufficient length of post streams. In addition, we randomized the stream of posts neglecting the timestamps. We found that the ob-

served stabilization is misleading. It is caused by the fact that the number of tags gathered in the profile grows with time and the impact of a single post on the distribution of profile tags decreases. When calculated in a sliding window, the fraction value does not seem to be related to the size of the profile (Fig. 3(c)). It suggests a weak potential impact of the collaborative model, as there is no relation between the tagging behaviour and increasing popularity of a resource (begin and end of a post stream). In addition, for a pair of tags that convey the same meaning, we would expect the collaborative effort of users to pick the dominant form of a tag and stop using the other one. To examine this, we selected a set of post streams from the *Delicious* dataset with at least 200 posts. We calculated the difference in disproportion between the two forms of the same tag after 100 and 200 posts, $\Delta = |0.5 - sf_{200}| - |0.5 - sf_{100}|$. The average difference calculated for the post stream as well as its randomized version is negligible (equal to -0.00766 and -0.00681 respectively). We found no evidence that collaborative behaviour of users leads to increasing preference of the dominant tag form.

HYPOTHESIS 3.2. *The shared knowledge model can explain the tag form popularity bias and constant use of both forms of the tag.*

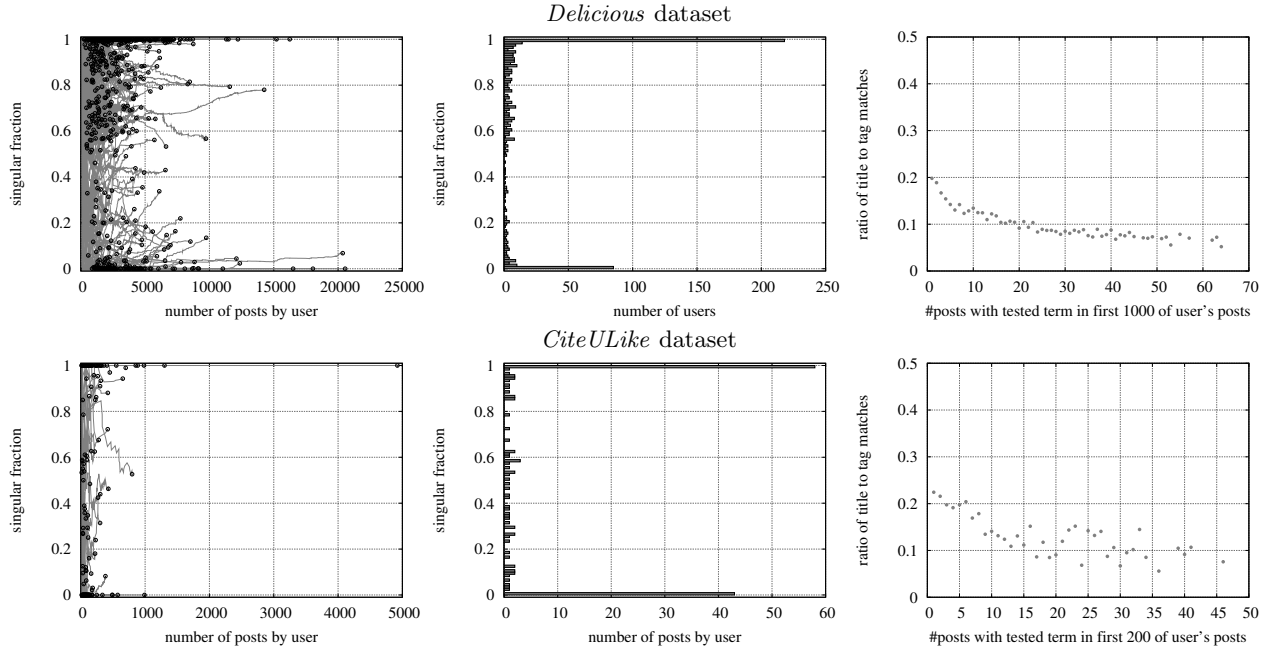
To test the potential impact of shared knowledge model, we examined the difference in frequency of occurrence for terms in each (singular, plural) pair in tag distribution over all posts and compared it to the difference in title words distribution over all resources. It was possible for the *Delicious* dataset only as we did not have the full information about title word frequencies for *CiteULike* dataset. As tags used in the test are commonly used words they could be frequently found in titles. The frequency of occurrence of title words used in the experiments was between 307 and 260,316. We represented each tested pair as a point in a two dimensional space. The x coordinate represented the difference between the frequency of singular and plural form of a term used as tags and the y coordinate represented the difference in frequency of using both term forms as title words. *Pearson's correlation coefficient* equal to 0.60 indicates strong correlation between tags and titles. Some terms may simply "sound better" in singular (or plural) form hence they are used more frequently in this form both in the title and tags.

4.2.2 The impact of resource title on user profile

HYPOTHESIS 3.3. *The personal model can explain the tag form popularity bias and constant use of both forms of the tag.*

To test the potential impact of personal model, we ran the previous experiment focusing on profiles of users, not resources. We picked users who used one or both forms of a tag frequently (at least 50 times for *Delicious* dataset and 10 times for *CiteULike* dataset). This time we were not able to classify the post stream traces based on the occurrence of one of the forms of the term in resource title, because users tag various resources. However, even neglecting the occurrence of the term in the title, the traces of user profiles lead to interesting observations. Most of the users pick a single form of a tag and use it consistently every time they tag a resource related to the concept represented by this tag. As most of the user profile traces have extreme values of *singular fraction* they overlap on the trace plot (Fig. 4(a)). To make this fact clear we present a histogram of final values of *singular fraction* for each user post stream (Fig. 4(b)). The histograms for two example pairs of tags (*blog(s)* and *network(s)*) show that the majority of users use the form of a tag, which generally is more popular, but a large group of users uses the other form only. Such behaviour is likely one of the factors that keep the constant inflow of both forms of a tag to the resource profile. However, at the same time this observation seems to contradict the results of the previous experiments. Most of the users are likely to completely disregard any external influence, including the title, as they have already decided on the tag that is going to represent a concept throughout their posts. It is important to notice that this experiment illuminated the behaviour of a specific group of users, who used the interesting tag frequently. Such frequent tags could be of special interest to the users as defining their general area of interests. We could imagine another group of users who used the same tag infrequently. For them the tag is most likely just an additional tag which only specifies the description of the resource.

Our hypothesis was that drawing a tag from the title is more likely for "infrequent" users than "frequent" users. To confirm this hypothesis we picked a set of users with large profiles ($p_{user} > N$, where $N = 1000$ for *Delicious* dataset and $N = 200$ for *CiteULike* dataset) who used at least one tag from the list of synonymous tags (192 tags for *Delicious* dataset and 102 tags for *CiteULike* dataset). To eliminate the bias caused by different sizes of user profiles we limited them to the first N posts entered by the user. The users were chosen separately for each of the traced tags. For each user/tag pair we recorded how many times k the tag was used among the first N posts of the user. Later, for each tag and each value of k we checked whether the use of the tag co-occurred with the occurrence of the term as title word. This allowed us to calculate the ratio of title to tag matches – the number of times the tag was used and it appeared in the resource title – to the total number of times the tag was used. To avoid the need for arbitrary choice of the threshold of k that would separate "infrequent" and "frequent" users we aggregated the results for each value of k separately. Despite the high variability of results for high k, some correlation between the frequency of tag use k and co-occurrence of title words and tags can be observed. The *Pearson's correlation coefficient* between k and the ratio of title to tag matches is equal to -0.21 for *Delicious* dataset and -0.25 for *CiteULike* dataset. High variability of results, which affected the value of correlation coefficient, was caused by problems with finding a representative set of users who used the tag a specific number of times k, when k is high. In most cases, for high k the ratio of title to tag matches could be calculated based on the information from a single user/tag pair which makes the results noisy. To reduce the noise, we combined the results for all tested tags and discarded results for k if the number of users, for which we recorded the data, was lower than 10. Despite the fact that this procedure limited the maximal value of k, for which we had any information, it reduced the noise and revealed the pattern of decreasing ratio of title to tag matches with growing k (Fig. 4(c)). The probability of a tag being drawn from the title by the user decreases with the number of times the tag was used by this user. Hence, when choosing a tag, "infrequent" users are more likely to be influenced by the title than "frequent" users.



(a) *singular fraction* for frequent users of tags blog(s) (*Delicious*) and tags network(s) (*CiteULike*), circle marks the fraction for full profile

(b) *singular fraction* value histogram for full profiles of frequent users

(c) "ratio of title to tag matches" – the number of times the tag was used and it appeared in the resource title to the total number of times the tag was used

Figure 4: Results for user profile tracing and the percentage for tag-title matches in relation to the number of occurrences of a tag in user profile.

5. CONCLUSIONS AND FUTURE WORK

The results of a sequence of experiments paint a complex picture of the sources of influence on the tagging process. We showed a strong relation between the resource title and the choice of tags used to describe this resource. Although part of this relation is certainly caused by the limited vocabulary and knowledge model shared between users of collaborative tagging systems and authors of tagged resources, by tracking synonymic pairs of tags we were able to isolate and show the direct impact that a form of a word in the title makes on the choice of the term form used as a tag. It seems that user convenience and shared knowledge model are stronger factors than collaborative behaviour for tagging decisions, for which we hardly found any evidence. We also identified another significant factor that can influence users to favour one of the forms – consistency of user profile. Most of the users would choose to keep the same form of the tag throughout their profile, if the tag represents an important concept. Users are much more willing to be influenced by the title if they are not planning to use the tag frequently. We believe that this picture can contribute to a better understanding of the behaviour of users of collaborative tagging systems. The conclusions drawn from this work can have direct implications on folksonomy modelling. In addition, the presented results impact two related tasks – tag recommendation and automatic taxonomy extraction, which are the objectives of our future work. For tag recommendation they suggest that although title is a good source of tag recommendations, which was already shown in tag recommendation competitions [8], the recommendations from the title should be considered in relation to user profile tags. If user's profile contains tags with identical or similar meaning, these tags are likely to be chosen instead of title recommendation. The conclusions weaken the idea of folksonomy as a folk taxonomy. Title based influence on tag choices, lack of collaborative behaviour that would eliminate a particular form of a tag and strong interests of users in organization of personal repository makes the idea of folksonomy as self-organizing taxonomy questionable. The choice of a form of a tag is very dependent on the character of this tag in the user profile. Depending on the user, the same tag can be used frequently to represent a general concept or infrequently just to specify such concept. It suggests that potential taxonomic relations can be found within a single user profile; however, they can be contradictory across user profiles, which makes the automatic extraction of general taxonomies hard, if not impossible.

The additional data gathered during the study to extend the used datasets, tables with complete results and code used to run the experiments are available at:
`http://www.cs.dal.ca/~lipczak/titleImpact.php`

6. ACKNOWLEDGMENTS

We would like to thank Tom Crecelius and Robert Wetzker as well as the administrators of BibSonomy and CiteULike for making their datasets available to us. The research was funded by the Natural Sciences and Engineering Research Council of Canada and the MITACS NCE.

7. REFERENCES

[1] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW '07: Proc. the 16th international conference on World Wide Web*, pages 835–844, New York, NY, USA, 2007. ACM.

[2] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum. Exploiting social relations for query expansion and result ranking. In *Data Engineering for Blogs, Social Media, and Web 2.0, ICDE 2008 Workshops*, pages 501–506, 2008.

[3] D. Bollen and H. Halpin. The role of tag suggestions in folksonomies. In *HT '09: Proc. the 20th ACM conference on Hypertext and hypermedia*, pages 359–360, New York, NY, USA, 2009. ACM.

[4] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81(2):591–646, May 2009.

[5] C. Cattuto. Semiotic dynamics in online social communities. *The European Physical Journal C - Particles and Fields*, 46(0):33–37, Aug. 2006.

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

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

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

[9] F. Figueiredo, F. Belém, H. Pinto, J. Almeida, M. Gonçalves, D. Fernandes, E. Moura, and M. Cristo. Evidence of quality of textual features on the web 2.0. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 909–918, New York, NY, USA, 2009. ACM.

[10] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

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

[12] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM '08: Proc. the international conference on Web search and web data mining*, pages 195–206, New York, NY, USA, 2008. ACM.

[13] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *SIGIR '08: Proc. the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[14] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. BibSonomy: A social bookmark and publication sharing system. In *Proc. the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures*, 2006.

[15] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Trend detection in folksonomies. *Semantic Multimedia*, pages 56–70, 2006.

[16] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, Feb. 1912.

[17] S. Ju and K.-B. Hwang. A weighting scheme for tag recommendation in social bookmarking systems. In *Proc. the ECML/PKDD 2009 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2009.

[18] B. Krause, R. Jäschke, A. Hotho, and G. Stumme. Logsonomy - social information retrieval with logdata. In *HT '08: Proc. the nineteenth ACM conference on Hypertext and hypermedia*, pages 157–166, New York, NY, USA, 2008. ACM.

[19] M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag sources for recommendation in collaborative tagging systems. In *Proc. the ECML/PKDD 2009 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2009.

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

[21] M. G. Noll, C. A. Yeung, N. Gibbins, C. Meinel, and N. Shadbolt. Telling experts from spammers: expertise ranking in folksonomies. In *SIGIR '09: Proc. the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 612–619, New York, NY, USA, 2009. ACM.

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

[23] 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: Proc. the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190, New York, NY, USA, 2006. ACM.

[24] M. Tatu, M. Srikanth, and T. D'Silva. Rsdc'08: Tag recommendations using bookmark content. In *Proc. the ECML/PKDD 2008 Discovery Challenge Workshop, part of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 96–107, 2008.

[25] R. Wetzker, C. Zimmermann, and C. Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings*, pages 26–30. ECAI 2008, July 2008.

[26] R. Wetzker, C. Zimmermann, C. Bauckhage, and S. Albayrak. I tag, you tag: Translating tags for advanced user models. In *WSDM '10: Proc. the Third ACM International Conference on Web Search and Data Mining*, pages 71–80, New York, NY, USA, 2010. ACM.