Sharing Vocabularies: Tag Usage in CiteULike

Sylvie Noël
Communications Research Centre
3701, Carling Ave.
Ottawa, ON, Canada
1-613-990-4675
sylvie.noel@crc.ca

Russell Beale
School of Computer Science
University of Birmingham
Edgbaston, Birmingham, UK
+44 (0) 121 414 3729
r.beale@cs.bham.ac.uk

ABSTRACT

CiteULike is a collaborative tagging web site which lets users enter academic references into a database and describe these references using tags (categorizations of their own choosing). We looked at the tagging behavior of people who were describing four frequently entered references. We found that while people tend to agree on a few select tags, people also tend to use many variants of these tags. This lack of consensus means that the collaborative aspect of tagging is not as strong as may have been suggested in the past.

Categories and Subject Descriptors

H.5.3 [**Information Systems**]: Group and Organization Interfaces – *web-based interaction.*

General Terms

Human Factors, Performance.

Keywords

tagging, social software, collaborative tagging, folksonomy, CiteULike, web.

1. INTRODUCTION

Social websites that use tagging are often described as collaborative because people can use these sites to share content actively with others. The best known of these are surely del.icio.us (http://del.icio.us) – henceforth Delicious - for URLs, and Flickr (http://www.flickr.com) for photos. The immense popularity of these two web sites show that they fill an important niche for web users. Delicious has just recently reached the two millionth user signup [1]). It is harder to find recent figures for Flickr; in April 2005, Stewart Butterfield claimed that there were approximately 270,000 Flickr users, while in June 2005 (after Yahoo's acquisition of Flickr), the number had suddenly increased to 775,000 registered users [5].

However, just because people are using these social sites does not mean that they are using them collaboratively. Williams [11] notes that not everyone who visits websites like Flickr or Delicious are active participants; they look at the content entered by other users, but do not enter new content themselves. Even those who participate actively may not be using tags socially, or at least not in the most efficient way to share tags with other users.

Published by the British Computer Society

This is what Rader and Wash [10] concluded when they looked at how people tag in Delicious. They analyzed a sample of 349 URLs that appeared in Delicious' "popular" and "recent" pages; these pages were bookmarked by 10 to 500 users and were described with 10 to 200 tags. They find an interuser tag agreement of 0.17 averaged over the sample; in other words, random pairs of users chose the same tags 17 percent of the time.. The probability that a user entering a single tag will successfully find a specific web page varies between 0.08 and 0.19, depending on how the calculation is weighted. These results suggest that people do not agree on the terms to use to describe specific URLs.

On the other hand, Golder and Huberman [4], also looking at Delicious, believe that the data shows that, over time, people tend to agree on the same tags to describe web sites. They looked at 212 bookmarked pages and at 229 users. Looking at URLs, they found that, in most cases, after approximately 100 bookmarks, tag frequency becomes stable. Golder and Huberman believe that this shows that even if there are many unique tags used to describe a web site people still tend to coalesce around a few specific tags.

At first glance, these results appear somewhat contradictory. It may be that the differences in the two studies are the result of sampling different URLs. Or it may be that the low interuser tag agreement found by Rader and Wash [10] simply reflects that a large number of unique tags can drown out the tendency of users to agree on a few tags to describe a URL (as found by Golder and Huberman[4]). We decided to compare the two methods to measure tag agreement on a single data set in order to clarify what might be occuring.

2. CITEULIKE

CiteULike (www.citeulike.org) is an academic article social tagging site. It is meant to allow sharing of academic articles between users. It is free to register and to use. There are three ways to enter a reference into CiteULike: by bookmarklet, by hand, or by uploading a bibTeX file. The bookmarklet and uploading methods will try to extract the bibliographic information automatically. With the bookmarklet and manual methods the user can enter tags. A list of tags the user has previously created is presented to the right of the bibliographic entry section, and the user can click on the tags to add them to the list, or can type new tags. If a person does not enter any tags, the system automatically enters the "no-tag" tag.

When a person reaches the CiteULike site, they are presented with some of the recent references added to the online database as well as a tag cloud of the most active tags. In a tag cloud, the size of the tags reflects their popularity: the larger the font, the more often that tag is used. People can search the site in many ways: by looking at other users' reference lists, by looking at tags, by groups or by subject. Groups are collections of users who share something in common: they can be members of the same lab, or people interested in the same research subject. Subjects are a relatively new development in CiteULike. It is now possible to categorize papers according to their scientific subject; existing subjects include computer science, biological science, social science, medicine, engineering, economics and business, arts and humanities, mathematics, physics, chemistry, philosophy, and earth and environmental science.

CiteULike uses blind tagging [8]: when people add a new reference to their database, they see only their own tags, not the tags that other people might have added to that reference previously. This makes it different from Delicious, where users can see not only the list of their own tags, but also popular tags which other users have selected for a URL. In fact, Delicious goes one step further by suggesting tags for the user (suggestive tagging). Flickr is more similar to CiteULike: since people are uploading their own photos, the system cannot really suggest a tag. CiteULike has also been studied by Farooq et al. [2] who use their analysis to devise a set of metrics for tagging sites.

3. ANALYSIS

3.1 Data

Data was collated from the 4th of November, 2004 (a few days before the site went officially live on the 11th) to the 27th of April, 2006. Each line in the data file contains an article ID number, an anonymized version of the user ID, the entry's date and time, and a single tag associated with the article. If a person added more than one tag to an article the data is reproduced in a second line with only the tag information changing. During this period of time 6,985 people used the site to enter 199,512 different articles and 51,079 different tags were used to describe these articles.

3.2 CiteULike versus Delicious

CiteULike differs in some ways from Delicious. When a person enters a URL, Delicious suggests tags that other people have associated with that URL (suggestive tagging), while CiteULike only shows users their own tags (blind tagging) [8]. When people can see how others are tagging artifacts, this may influence their behavior such that they tag more for others than for themselves[7]. This would make entering an artifact in CiteULike cognitively different from entering one in Delicious: while the goal in the former is focused on personal retrieval at a later date, the goal in the latter is to share the artifact with other users. If this were the case, one would expect artifacts in CiteULike to show more heterogeneous tags and those in Delicious to show more homogeneous tags.

The user population of CiteULike is much smaller than that of Delicious. This means that there may not have been enough data entered to see a consensus form around tag choice for references. Nevertheless, we were curious to see whether our user population acted in a similar fashion to the users of Delicious.

3.3 Tag usage in popular articles

We wanted to see if we might find a combination of the low inter-user agreement on tags found in Rader and Wash [10] and the rallying around certain tags found in Golder and Huberman [4].

We looked at the four most popular articles in CiteULIke during the period covered by our data file, i.e. the ones that were referenced by the most people. Table 1 shows the number of users and tags associated with each of these articles. Two of these articles are about social bookmarking and tagging, one was about semantic blogging, and one about blogs and wikis.

Table 1. Articles used in this study

Article ID	Total number of users	Total number of tags
A	63	74
B	54	62
C	49	38
D	47	33

We calculated tag proportion in the following manner. For each paper, we counted each addition of tags by a user on a specific date as a separate point in time (so, if there were two people adding tags on the same date, we counted them as two different points in time). We then calculated the proportion of each tag at point X in time according to the total number of tags that had been added up to that point.

We also calculated the interuser agreement (IUA) in the following manner ([3][10]). For each pair of users, we counted the number of identical tags that were applied by both to a reference, and divided that by the total number of individual tags they used for that reference. We then calculated an average IUA for each article. IUA gives us another view of how much people agree on the tags used to describe references.

3.3.1 *Article A*

This paper is concerned with tags and folksonomies. It was first referenced on the 27th of August 2005 and was regularly added to CiteULike until the 26th of April 2006. As seen in Table 1, 63 users added this paper to their CiteULike database using a total of 74 different tags.

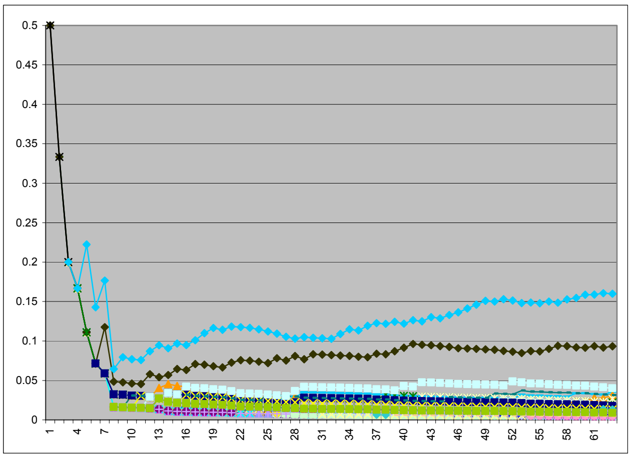

Figure 1. The relative proportion of tags for article A. Each line represents a single tag. The horizontal axis denotes time in units of article added to a user's library, the vertical time the proportion of use relative to all tags.

Figure 1 shows the relative proportion of tags used for this article. This figure resembles those seen in Golder and Huberman's paper, with a relatively quick stabilization of tag proportion at around the 50th entry. There is some agreement on two tags for this paper ("tagging" and "folksonomy"). However, most tags used only appear a maximum of three times. Indeed, the agreement on the two major tags is not very large, with the most popular tag only making up about 15 percent of all tags. There are slight variations that might have boosted the number for these tags (e.g., "folksonomies", "tag", "tags", "socialtagging") which show that people tended to agree on the article's subject, but that their vocabulary was not standardized[10]. In fact, the average IUA for this article is 0.13, meaning that random pairs of users chose the same tags 13 percent of the time. This number is halfway between the 8% reported by Furnas et al. [3] and the 17% reported by Rader and Wash [10].

3.3.2 *Article B*

This is an article on social bookmarking tools which first appears on the 17th of April 2005 and makes its last appearance in our data file on the 21st of April 2006. As noted in Table 1, 54 users added this reference to their CiteULike library, using a total of 62 different tags.

Figure 2 shows the proportion of tags used for this article. At the moment when the entries cease, stabilization does not yet appear to have been achieved, nor are there any clearly dominant tags for this paper. The tags most used are "tagging", "folksonomy", "bookmarking", "socialbookmarking" and "social-bookmarking". Each of these tags also has many variants appearing in the database. Again, this shows that people tend to agree on what the paper is about, just not what words to use to describe it. The IUA for this article was a low 0.07.

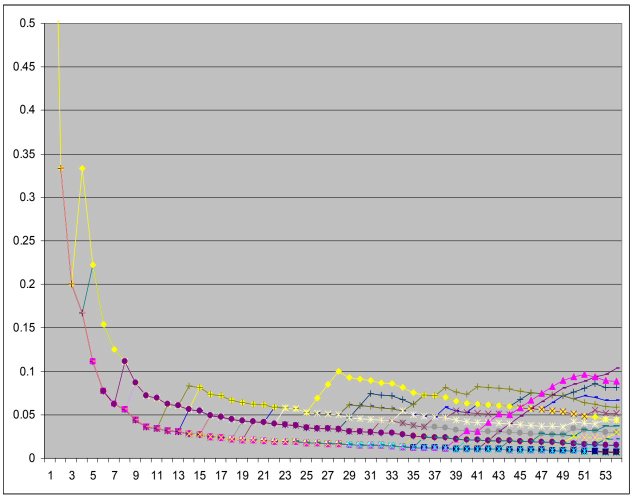

Figure 2. The relative proportion of tags for article B. See Figure 1 for explanation.

3.3.3 *Article C*

This is a paper on semantic blogging as a tool for knowledge management. It was first referenced on the 23rd of November 2004 and last appeared in our database on the 3rd of November 2006. As shown in Table 1, it was added by 49 users who used 38 different tags to describe it.

Figure 3 shows the proportion of tags used for this article. Stabilization appears to occur around the 45th entry, with people rallying around four popular tags ("blogging", "semantic", "km", and "blog"). Again, there are many variants for each of these tags; for example, "blogging" could be associated with "blog", "bloging" (*sic*), "blogs", "weblog", and "weblogs". Adding all these tags up actually doubles the number of people who describe this paper as having to do with blogging (from 14 to 33). For this article, the IUA was also low, at 0.09.

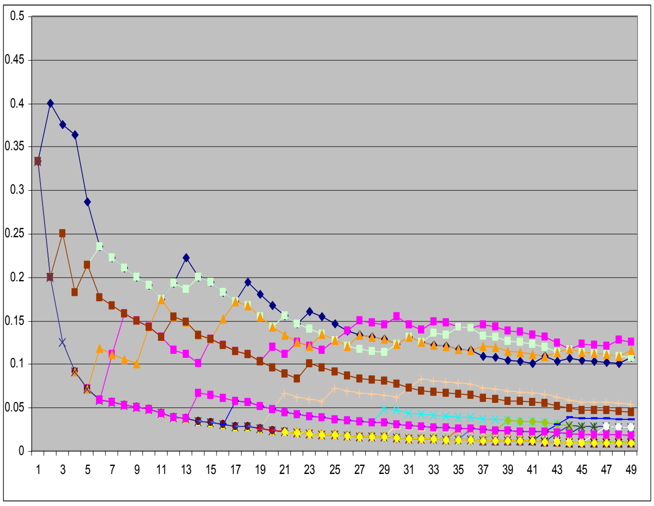

Figure 3. The relative proportion of tags for article C. Each line represents a single tag. See Figure 1 for explanation.

3.3.4 *Article D*

This paper is about blogs and wikis. It was first added to CiteULike on the 11th of January 2005 and made its last appearance in our database on the 16th of March 2006.

Figure 4 shows the relative proportion of tags used for this article. Again, we see a fairly quick stabilization of tag usage by about the 40th addition. Three tags stand out for this paper ("wiki", "blogs" and "blog"). If we put aside spelling, people mainly describe this paper as having to do with weblogs ("blog", "blogging", "bloging" (*sic*), "blogs", "weblog", and "weblogs") and wikis ("wiki", "wikis"). The IUA was similar to the first article, at 0.16.

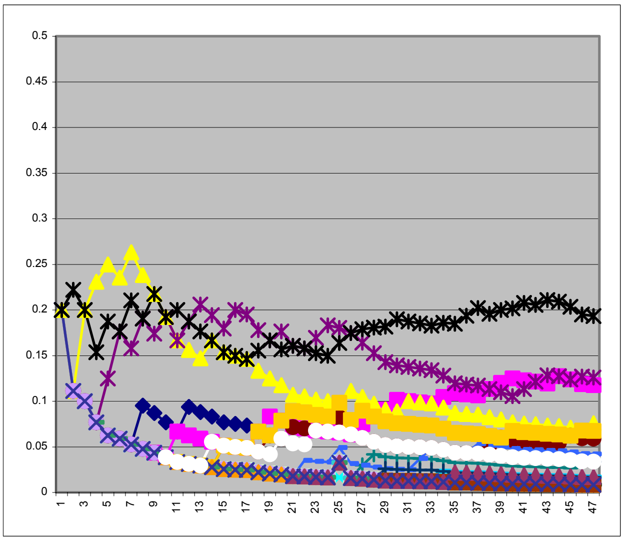

Figure 4. The relative proportion of tags for article D. Each line represents a single tag. See Figure 1 for explanation.

4. Discussion

Our results are similar to those obtained by Golder and Huberman [4]: in three out of four cases there is a fairly quick stabilization of tag usage, indeed much quicker than found by Golder and Huberman. We also find low interuser agreement for the same data: IUA varies from 0.07 to 0.16. These numbers are similar to those found by Rader and Wash [10]. It would seem then that the results from these two research groups are not contradictory: while an important number of users agree on a few tags to describe a specific artifact, most tags for that artifact are used by only a few users.

Golder and Huberman suggest that stabilization might occur because of imitation or shared knowledge. Because CiteULike does not propose tags, imitation is probably not the source of the tag stabilization seen for its references (it is theoretically possible that people, when adding a reference, will check to see if that reference already exists in CiteULike and how people have tagged it; but this is complicated to do and we doubt that users would proceed this way). This leaves shared knowledge as an explanation. This is most probably the source of the stabilization. Academic papers are usually on very specific subjects and the people using CiteULike are, usually, working in the domain referenced by the academic papers they enter and therefore share a common vocabulary with their colleagues who enter the same paper. Once enough tags are entered, a social consensus begins to take shape in the data, even when tag entry is blind to other users' choices.

Golder and Huberman ([4], p. 206) state that "(t)he commonly used tags, which are more general, have higher proportions, and the varied, personally oriented tags that users may use can coexist with them". We did not find that 'commonly used tags' were more general than the low-proportion tags. Many of these were in fact variations on the most common tags (plural/singular, alternate spelling, etc.). This "vocabulary problem" is well known [6, 9].

In summary, while a few tags prove a bit more popular than the rest, for the most part people tend not to use the same words when tagging a reference, even when they are thinking of the same concept, and even in specialist domains.

5. ACKNOWLEDGMENTS

The authors would like to thank Richard Cameron both for creating CiteULike in the first place, and in providing them with the data for this study.

6. REFERENCES

[1] Conradh, Del.icio.us blog, *That was fast*, entry for March 29, 2007, available at http://blog.del.icio.us/blog/2007/03/that_was_fast.html.

[2] Farooq, U., Kannampallil, T., Song, Y., Ganoe, C., Carroll, J. and Giles, L., Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. in Proceedings of the 2007 international ACM conference on Conference on supporting group work, (2007), ACM New York, NY, USA, 351-360.

[3] Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T. Statistical semantics: Analysis of the potential of key-word information systems. *The Bell System Technical Journal*, vol. 62 (6), 1753-1806, 1983.

[4] Golder, S.A. and Huberman, B.A. Usage patterns of collaborative tagging systems. *Journal of Information Science*, vol. 32 (2), 198-208, 2006.

[5] Google Answers, *Flickr User Base*, 22 Julay 2005, available at http://answers.google.com/answers/threadview?id=546695

[6] Heyman, P. and Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems, *Stanford Infolab Technical Report*, No. 2006-10. 24 April 2006.

[7] Lawley, E.L. *Social consequences of social tagging*, 2005, available at http://many.corante.com/archives/2005/01/20/social_consequences_of_social_tagging.php

[8] Marlow, C., Naaman, M., boyd, d., and Davis, M. HT06, tagging paper, taxonomy, Flickr, academic article to read in *ACM Hypertext 2006 – Seventeeth ACM Conference on Hypertext and Hypermedia*, Odense, Denmark, ACM Press, 2006.

[9] Mathes, A. *Folksonomies – Cooperative Classification and Communication through Shared Metadata*. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.pdf

[10] Rader, E. and Wash, R. Tagging with del.icio.us: Social or selfish? *Conference on Computer Supported Cooperative Work, Conference Supplement*, pp. 211-212, Banff, Canada, ACM Press, November 4-8, 2006.

[11] Williams, I. Users failing to interact with Web 2.0 sites *Infomatics*, 20 April 2007, available at http://www.infomaticsonline.co.uk/vnunet/news/2188259/users-really-interacting-web