Tagging with del.icio.us: Social or Selfish?

Author
Affiliation
Location
Email address

Author
Affiliation
Location
Email address

1. INTRODUCTION

del.icio.us is a website for "social bookmarking" where users can store and access their bookmarks online, along with descriptive keywords or "tags." When a user of del.icio.us logs in to their account and adds a bookmark, she may also tag that bookmark with any 10 or fewer single words that she feels are somehow related to that web page. Both the tags and the bookmarks are then publicly available; searching by a tag produces all of the bookmarked web pages ever tagged with that word. Because the tags are public, it is possible that users' choices regarding what tags to apply could be influenced by the tagging practices of others, and a consensus might emerge for which tags should be used in a given context [6]. However, it has long been accepted that people use language imprecisely, and meaning is negotiated on-the-fly during conversation [2]. This imprecision is evident not only in communication, but also when people are asked to create keywords for recipes and names for common editing operations [4], and when user-generated index terms are compared with Library of Congress subject headings [3]. In fact, the probability that two people will generate the same label for the same object is widely held to be less than 20% [1,4].

For personal bookmarking and sharing with friends, the "vocabulary problem" isn't much of one; "who" and "how many" people saved the bookmark are more salient. But when a user wants to take advantage of the collective benefit of social tagging by browsing or searching tags, the problem is more apparent. Users who tag selfishly, without thinking about the public audience for their tags are unlikely to select the same tags as other users. In del.icio.us this diversity enhances findability; unfortunately at the expense of community convergence on a recognized and learnable tag vocabulary. When tags are used inconsistently, the user can expect more seemingly irrelevant information to be returned for each tag. A tradeoff exists: social users of tagging are better off if there is convergence, while findability is improved if users tag selfishly.

A question remains about whether users of del.icio.us practice social or selfish tagging. Marlow et al. [7] points out that while some people use tags for the purpose of organizing their own bookmarks, others intentionally choose to contribute to the value of "conceptual clusters", or call attention to the pages they bookmark, by adhering to conventions. Golder & Huberman [6] report that over time the relative frequencies of tags applied to a web page stabilize into a pattern such that the most commonly used tags remain so and do not fall out of favor. They speculate that this could be because users imitate each other, or the user community is similar enough to naturally tag things the same way, or the stable content of the web page itself acts as a limit on tags people might choose.

If it is true that in general users of del.icio.us actively practice social tagging, then it is reasonable to assume that an analysis of tag frequencies for individual users and web pages would NOT show a similar pattern to the original "vocabulary problem" work published by Furnas et al. [4]. Therefore, an analysis of bookmark, user and tag data for 349 web pages downloaded via del.icio.us was conducted to discover whether the "vocabulary problem" is present in the way users select tags for web pages. Results indicate that there is very little inter-user agreement, suggesting that most users consciously or inadvertently tag selfishly. These tagging practices have important implications for the findability of web pages in del.icio.us.

2. DATASET

We collected a sample of 500 web pages that were bookmarked on del.icio.us, and listed on the "popular" and "recent" pages on several days between August and November 2005. We downloaded the "URL pages" for each, which list every user who bookmarked the web page and all the tags they applied. All "URL pages" were then re-collected at one time, in late December 2005, to obtain a consistent snapshot in time, and parsed to pull out the relevant data. Tags in del.icio.us are not case sensitive, but the system is sensitive to misspellings, tenses, and plurals. So for example, we treated the words "book" and "books" as unique tags, but "book" and "BOOK" as equivalent. Then, we eliminated the "extreme" web pages from the sample, retaining those bookmarked by 10 to 500 users (20^{th} to 80^{th} percentiles), and with 10 to 200 tags (20^{th} to 95^{th} percentiles). Both of these variables exhibit long-tail distributions consistent with Zipf's law [1,4] and we believe the 349 web pages that remain represent "typical" usage patterns for that time period.

We replicated analyses from Furnas et al. [4] to determine whether the vocabulary problem exists in del.icio.us. Because del.icio.us stores every tag (word) that is used to refer to every bookmarked web page (object), we chose a set of analyses from [4] that most closely approximated these parameters, called "Several names per object". We used two measures that estimate in different ways the "repeat rate", or likelihood that a tag generated by a user is among the tags the system already has stored for that web page. Repeat rate indicates how likely it is for a search on a single tag to succeed. Finally, we

Conference'04, Month 1–2, 2004, City, State, Country.

also calculated inter-user tag agreement for all users who had bookmarked the 349 web pages [3]. This measure tells us on average, how often random pairs of users generate the same tag for the same web page.

3. RESULTS

The 349 web pages in our sample were bookmarked by 120.23 users (σ = 111.25) who used 3.02 tags apiece (σ = 0.58), on average. These web pages also tended on average to have 378.48 tags associated with them in total, 57.21 of which were unique. A total of 21,976 users were involved in bookmarking one or more of the 379 web pages, using 9557 unique tags.

The three most common tags for each web page represented 45% of all tags applied for an average page, which is greater than the 33% reported by Furnas et al. [4]. Inter-user tag agreement, averaged over the sample, was 0.17, meaning that random pairs of users chose the same tag for the same web page just 17% of the time (σ = 0.10). While this percentage is low, it is higher than the 8% reported in [4] for their text-editing operations dataset.

Repeat rate statistics were calculated in two ways. In the first calculation, "weighted random", when a user searches for a single tag the probability of success depends upon the relative frequencies with which different tags are associated with web pages. Pages that are associated more frequently will be returned more often than others. The second calculation, called "optimized" in [4], rank-orders the web pages for each tag by frequency, and always returns them such that the highest frequency page is returned first, then the next highest, and so on. Table 1 shows comparison of the repeat rate statistics reported in [4] with those calculated for our sample. The letter M represents the number of tags the system stores, per web page.

Table 1. Comparison of repeat rate statistics

	Furnas [4] Common objects	del.icio.us weighted random	del.icio.us optimized lower	del.icio.us optimized upper
M=1	.12	.08	.17	.19
M=2	.21	.15	.28	.31
M=3	.28	.22	.37	.37

The values in Table 1 represent the probability that a user entering a single tag will be successful in their search for a specific web page, depending on how many tags are stored in the system for each web page. Success rates are very similar between our sample and the results reported in Furnas et al. Figure 2, below, illustrates the optimized lower and upper bound values for M=1 through M=40. Success rate increases dramatically between 1 and 10 tags (words) for our sample, and then appears to level off. In [4], success rate appears to level off after about 8 words. These results lead to the conclusion that the "vocabulary problem" does exist, and that selfish tagging, not social, is prevalent. `del.icio.us` solves the vocabulary problem for searching for individual web pages, but not for situations where a user wants to learn more about a particular topic.

4. CONLCUSION

`del.icio.us` supports both social bookmarking and social tagging, two ways to interact with web pages stored by others. Social bookmarking depends upon nothing more than the public

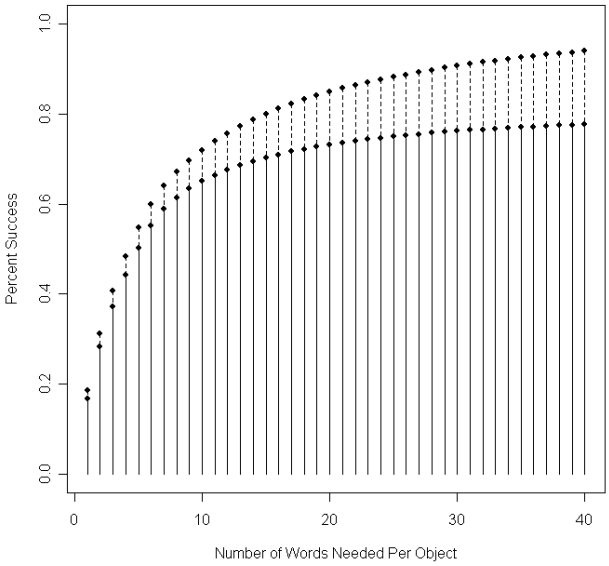

Figure 1. Percent success based on "weighted random", as a function of the number of words stored for each object.

nature of the bookmarks individuals store online. However, social tagging depends on the formation of conventions for how tags are applied to content. When tags are applied inconsistently, it becomes difficult for a user who wants to learn about a particular topic to sort out what she means when she uses the tag, from what others mean when they use it. `del.icio.us` is able to store a large number of tags for every object, which dramatically increases the probability that a search for an individual web page will be successful. However, this comes at the expense of the social aspects of tagging. One can imagine the possibility of a system in which both selfish and social tagging coexist with equal levels of success; however, this would likely require a human or algorithmic indexer, or editorial control over tag synonyms and usage [1]. The goal would be to eliminate inconsistency in the way common tags were used, while retaining the rare tags.

5. REFERENCES

[1] Bates, M. J. (1998). Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science, 49(13),* 1185-1205.

[2] Clark, H. H. (1996). Using Language. Cambridge, England, UK: Cambridge University Press.

[3] Collantes, L. Y. (1995). Degree of Agreement in Naming Objects and Concepts for Information Retrieval. *Journal of the American Society for Information Science, 46(2),* 116-132.

[4] Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1983). Statistical Semantics: Analysis of the Potential Performance of Key-Word Information Systems. *The Bell System Technical Journal, 62(6)*, 1753-1806.

[5] Golder, S., & Huberman, B. A. (2006). The Structure of Collaborative Tagging Systems. *Journal of Information Science, 32(2),* 198-208.

[6] Marlow, C., Naaman, M., boyd, d., & Davis, M. (2006). Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In the *Proceedings of WWW 2006 Collaborative Web Tagging Workshop*, Edinburgh, Scotland.