Understanding the Efficiency of Social Tagging Systems using Information Theory

Ed H. Chi, Todd Mytkowicz[+]
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304 USA
echi@parc.com, [+]Todd.Mytkowicz@colorado.edu

Abstract

Given the rise in popularity of social tagging systems, it seems only natural to ask how efficient is the organically evolved tagging vocabulary in describing any underlying document objects? Does this distributed process really provide a way to circumnavigate the traditional categorization problem with ontologies? We analyze a social tagging site, namely del.icio.us, with information theory in order to evaluate the efficiency of this social tagging site for encoding navigation paths to information sources.

Introduction

The accumulation of human knowledge relies on innovations in novel methods of organizing information. Subject indexes, ontologies, library catalogs, Dewey decimal systems are just a few examples of how curators and users of information environments have attempted to organize knowledge. Recently, tagging has exploded as a fad in information systems to categorize and cluster information objects (Furnas 2006). Tagging has become a useful way for users to recall information sources for later use as well as to communicate interesting nuggets of information to other users (Hammond 2005).

Some bloggers have written about the organic nature of the evolution of the tags in a social tagging system and how it compares with ontologies, and perhaps the most well-known writing is Clay Shirky's work (Shirky 2006). Shirky argues that since tagging systems does not use a controlled vocabulary, it can easily respond to changes in the consensus of how things should be classified. Indeed, Golder and Huberman (2006) studied the growth of tagging systems, and offered a potential explanation for why objects acquire stable tag patterns.

Furnas et al. (2006) pointed to the usefulness of social tagging systems as a communication device that can bridge the gap between document collections and users' mental maps of those collections. Social navigation as enabled by social tagging systems can be studied by how well the tags form a vocabulary to describe the contents being tagged.

Indeed, to understand how tags have evolved for a large corpus of tagged items such as del.icio.us, we need to understand whether the tags adequately describe the items being tagged. Moreover, we need a way to understand how the social tagging system will evolve in the future.

In short, our contribution is providing a methodology for understanding how tags in a social tagging system evolve as a vocabulary. How well does social tagging afford social navigation of a large collection of sources? We are proposing the use of concepts in information theory to examine this problem.

Method

A bookmark in a social tagging system can be viewed as a 3-tuple consisting of a unique identifier for the document object, a user, and a set of tags. Let D denote the set of documents, U users, and T tags. Let B denote a set of bookmarks. Then a single bookmark b is a single document d, a single user u, and a set of tags $t_1, \ldots, t_n$. Without loss of generality, it is then possible to express the bookmark b as a set of 3-tuples $(d, u, t_1) \ldots (d, u, t_n)$. In our data, we decompose all of the bookmarks into this form.

We collected del.icio.us bookmarking data using a custom web crawler and screen scraper. Our crawling tool walked the del.icio.us site and dumped the parsed bookmarks into a MySQL database for analysis. We started at the del.icio.us homepage and harvested a set of users. For each user, we collected their bookmarks, as well as links to other users that have bookmarked the same document. In essence, our crawler did a random walk of the graph over users and documents. Done over a 2 month period with some 40 machine, we are somewhat confident that our collection is a fairly-complete del.icio.us data up to late-summer 2006. We had 9,853,345, documents 140,182 users, and 118,456 users in our dataset.

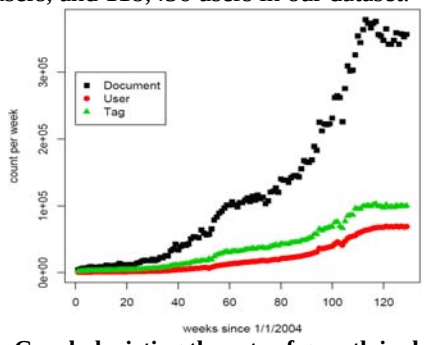

Figure 1. Graph depicting the rate of growth in documents, users, and tags over time. Plotted by count per week.

Results

We first computed the frequency and then the corresponding probability distribution of the documents being bookmarked in the del.icio.us. Using this distribution we then generated the entropy curve for the tag and document set.

The entropy of documents conditional on tags, *H(D|T)*, is increasing rapidly (see Figure 2). What this means is that, even after knowing completely the value of tags, the entropy of the document is still increasing. This measure gives us a method for analyzing how useful a set of tags is at describing a document set. The fact that this curve is strictly increasing suggests that the specificity of any given tag is decreasing.

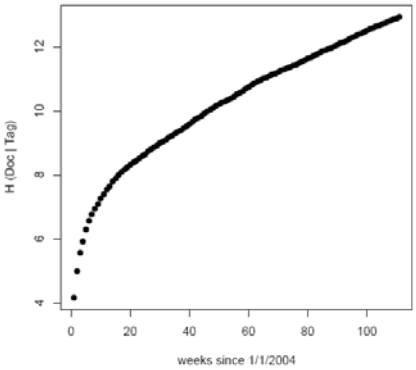

Figure 2. Entropy of Documents conditional on Tags *H(D|T)* increases over time.

The conditional entropy *H(T|D)* asks the reverse question of *H(D|T)* discussed in the previous section. "If I know a set of document, what uncertainty remains in the tags that are used to describe these documents?" Interestingly enough, *H(T|D)* has been increasing steadily as shown in Figure 3.

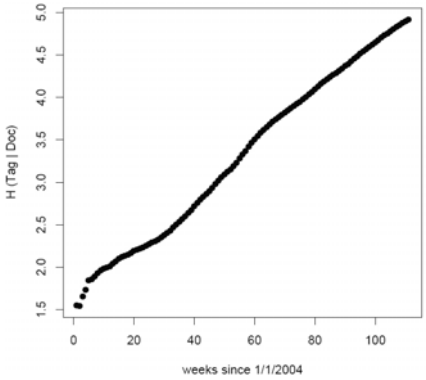

Figure 3. Conditional entropy *H(T|D)* increases over time.

Figure 4 illustrates the mutual-information *I(X;Y)* over time. Mutual information is a measure of independence. Full independence is reached when *I(X;Y) = 0*. As seen in Figure 4 the trend is steep and quickly decreasing.

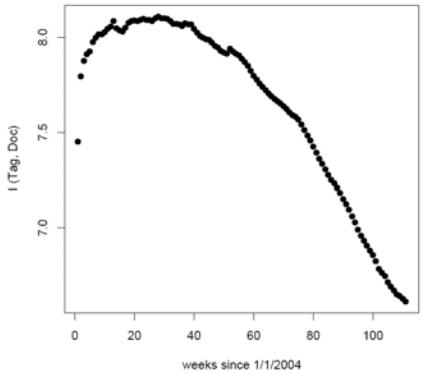

Figure 4. Mutual Information *I(T;D)* decreases over time showing that tags are becoming less descriptive for any given document in del.icio.us. A value of 0 means complete indepenence between Tags and the Documents those tags are supposed to encode.

Discussion and Conclusion

We seek to understand the efficiency of shared tagging vocabulary emerging from distributed social taggers on the web. We used information theoretic measures to try to quantify the complex dynamics in a social tagging site. The result suggests that the quality of the shared mapping between tags and documents in del.icio.us appears to be getting noisier over time.

In order to increase the effectiveness of the encoding process in del.icio.us, one needs to decrease *H(D|T)*. Given that *H(D)* is out of the control of the designer, all that can be controlled explicitly via the designer's of tagging software is *H(T)*. Current tagging interfaces usually provide ``popular tags" when any individual user attempts to encode a document. In effect, by providing this facility to ease the encoding process for the tagger, the designer's of tagging sites are causing *H(T)* to become less diverse. Rather than providing popular tags for user's, tagging sites should ask them to think of tags that describe the document that are not in the popular list.

References

Furnas, G. W., Fake, C., von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C., and Naaman, M. 2006. Why do tagging systems work?. In *Proc. CHI '06 Extended Abstracts*. ACM Press, New York, NY, 36-39.

Golder, S. and B. A. Huberman. (2006). Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2). 198-208.

Hammond, T., T. Hannay, B. Lund, and J. Scott. Social bookmarking tools : A general review. *D-Lib Magazine*, 11(4), April 2005.

Shirky, C. Ontology is Overrated: Categories, Links, and Tags. Blog entry. http://shirky.com/writings/ontology overrated.html (retrieved Sept 21, 2006).