Understanding Navigability of Social Tagging Systems

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 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 navigation to information sources. We show that over time, del.icio.us is becoming harder and harder to navigate and provide an evaluation metric, namely entropy, that can be used to evaluate and drive system design choices.

Author Keywords
Social tagging, navigation, information access, ontologies, efficiency, evaluation, methodology, information theory, entropy.

ACM Classification Keywords
H.5.3 [Information Interfaces and Presentation]: Group and Organization Interfaces—collaborative computing; H.1.2 [Models and Principles]: User/Machine Systems—Human information processing; K.4.3 [Computers and Society]: Organizational Impacts – Computer-supported collaborative work; H.5.2 [Information Interfaces and Presentation]: User Interfaces

General Terms: Design, Experimentation, Human Factors

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 [8]. 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 [11].

The purpose of an organization scheme for any information system is to help users browse, discover, and navigate through the information sources. A unique aspect of tagging systems is the freedom users have in choosing the vocabulary that can be used in tagging objects. First, tags are not required to be placed in a hierarchy. Second, any free-form keyword using a combination of letters and numbers is essentially allowed as a tag.

Surprisingly, the use of free-form labeling may appear like a recipe for disaster but instead has turned into one of the newest trends on the web. Indeed, social tagging has been applied to photos [6], videos [20], web pages [5], Wikipedia articles, and academic paper citations [4], and in each of one of these cases, an organically-evolved vocabulary has emerged to describe the content of the tagged objects. 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 [16].

An important question is raised by Shirky's essay, "How and why do social tagging systems work?" This is enough of a mystery such that a special panel during last year's CHI2006 conference brought in experts to discuss this topic [8]. Some research has also emerged to explain and characterize the tagging phenomenon [10, 13]. Golder and Huberman studies the growth of tagging systems, and offered a potential explanation for why objects acquire stable tag patterns [10]. Macgregor and McCulloch provide an overview of the phenomenon and explore reasons why both social tagging as well as ontologies will have a place in the future of information access [13]. From a system oriented perspective, Sen et al. studied how personal tendencies and community influences affect the way users tag items on a movie recommender website [15]. This set of initial research points to the fascination with social tagging.

During the CHI2006 panel, George Furnas mentioned that a potential cognitive process for explaining how social tagging works might arise out of an analysis of the "vocabulary problem" [8]. Specifically, Furnas mentioned that the process for generating a tag for an item that might

be needed later appears to be the same process that is used to generate search keywords to retrieve a particular item in a search and retrieval engine [7]. He also noted that generating a small set of keywords for an item is relatively easy for a single user, while generating a large set of keywords is a task best left for a group of users. In short, "different users use different terms to describe the same thing" [8].

Viewed from this perspective, social tagging and other indexing systems are attempting to solve a mapping problem. Social tagging, in a way, is trying to provide a map, which are summarizations of an explorable space. Using this map, users can navigate within the information space much more effectively than they otherwise would be able to do. The vocabulary problem in social tagging is thus asking the question of how groups of users effectively generate this map (a set of keywords mapped to a set of items).

Indeed, the raging debate between using ontologies vs. social tagging boils down to the fact that both approaches are trying to index the content for later retrieval. Thus, the relative efficiency in generating good maps could either rest within predefined ontology controlled by a rule system, or emerge organically amongst the wisdom of a crowd.

Therefore, our task in understanding how social tagging is evolving is reduced to the problem of understanding how social tagging affords social navigation. Given a tag vocabulary and its mapping to items, how effective is it in describe that set of items?

In this paper, we propose to use information theory to analyze the effectiveness of social tagging. We have applied this methodology to the entire del.icio.us social tagging systems. The results show that:

1) The efficiency of social tagging is decreasing. First, the entropy of the tag vocabulary is increasing. Moreover, the entropy of documents conditional on tags, *H(docs|tags)*, is also increasing. Together, these two pieces of evidence mean that tags are becoming less and less descriptive. These results suggest that tag effectiveness is decreasing, but from a design perspective, there are ways to increase tag effectiveness.

2) The efficiency of social navigation afforded by the tagging system is also decreasing. First, the entropy of the documents is increasing. Moreover, since *H(tags|docs)* is increasing, it is becoming harder for users to use the tag vocabulary to describe the document objects they want to bookmark. In other words, tags are becoming less meaningful in regards to providing salient navigability.

3) We found that the entropy of user *H(Users)* continues to increase, suggesting that there is a greater diversity of users utilizing the social bookmarking system. The entropy of documents conditional on users, *H(docs|users)*, was increasing but has plateaued. Combining that with the fact that the total number of users is increasing, the data strongly suggests that users are increasingly discovering similar sources of documents, and that they overlap in places they visit more strongly than before.

Finally, these analyses have direct design implications for social navigation tools. Not only do they provide a methodology whereby one can evaluate new system design, they also have strong implications for how people are using the web at large and provide motivation for new systems.

RELATED WORK

Social bookmarking systems have actually been a research area for some time, including systems such as Glance et al's Knowledge Pump [9], Bouthors and Dedieu's Pharos [2]. However, these systems did not use social tagging. The use of free-form labeling to tag documents has actually been around for a while, but became a focus of the Web 2.0 movement when Thomas Vander Wal described the phenomenon as "folksonomy" [18]. The phenomenon is now used to describe Web-based technology for generating free-form labels that categorize contents collaboratively. The popularity of these social tagging systems perhaps can be attributed to the benefits users perceive in easily recalling contents that might be useful later.

The surprising aspect of the phenomenon is that, in contrast to professionally developed taxnomies, a folksonomy appears rather unsystematic, but over time, an order within the tags appears. Some have posited that this is due to the fact that social tagging systems dramatically lower the cost of labeling items when compared to traditional taxonomies, because one does not have to be trained to use the taxonomy [13]. Since many users can do labeling cheaply and in a distributed fashion, many more objects are tagged. Moreover, since it does not use a controlled vocabulary, it can easily respond to changes in the consensus of how things should be classified. This is a point well-explored by Shirky in his essay [16].

There are just starting to be a handful of academic research focused on the dynamics of social tagging systems. The most well-known so-far is probably Golder and Huberman's initial work on understanding the usage patterns of social tagging systems [10]. They characterized a small subset of del.icio.us data and found that there is potentially a growth pattern to the tagging system. Moreover, they found that, for an item, tags slowly stabilize to a pattern in which the proportion of each tag is a fixed percentage of the total frequency of all tags used.

Library scientists have also started to look at social tagging and its relationship to traditional taxonomies. MacGreogor and McCulloch collected and discussed a set of arguments related to the pros and cons of controlled vocabulary vs. free-form labeling [13].

CSCW researchers have also started looking at the field as a area for investigation. In particular, Sen et al. studied how tag selections are affected by community influences and personal tendencies [15]. They studied four different tag selection algorithms for displaying tags from other users and found that user tagging behaviors changed depending on the algorithm.

Millen et al., on the other hand, introduced a system designed for tagging intranet items in an enterprise [14]. They collected some sample user data, and showed that users form social networks through patterns of their usage of others' bookmarks.

Within the HCI community, the CHI2006 panel organized by Furnas et al. brought social tagging systems to the attention of HCI practitioners [8]. Furnas' suggestion that social tagging systems can be viewed from a "vocabulary problem" [7] perspective directly inspired the approach used in this paper. More importantly, his comment 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.

MacGregor also refer to social tagging fundamentally as a vocabulary problem in indexing: "terms assigned to resources that are exhaustive will result in high recall at the expense of precision. Conversely, terms that are too specific will result in high precision, but lower recall.[13]" Sen et al. also referred to this as a fundamental research problem in the conclusion of the paper: "the density of tag applications across objects may provide information about their value to users. A tag that is applied to a very large proportion of items may be too general to be useful, while a tag that is applied very few times may be useless due to its obscurity.[15]"

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. Will the tags lose their specificity?

In short, our contribution is providing a methodology for understanding how tags in a social tagging system evolve as a vocabulary. And 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 these questions.

ENTROPY AND INFORMATION THEORY PRIMER

Information theory has long been applied to the understanding of human languages. Claude Shannon invented much of information theory in 1948 to understand how codes can be efficiently transmitted in a communication channel. The concept of source coding and channel coding in information theory is concerned with the efficiency and robustness of codes in communication. These concepts have been applied to human languages. Codes are essentially words, which should be efficient and short if they are commonly used, e.g. "I", "the", "a", "this". Second, a language (that is, codes used in communication) should be robust to noise by having enough redundancy in the system.

The central idea in Shannon's theory is the source-coding theorem, which says that, on average, the number of bits needed to communicate the result of an uncertain event is given by its entropy. From a statistical perspective, a somewhat more intuitive understanding of *entropy* is that it measures the amount of *uncertainty* about a particular event associated with a probability distribution.

The concept can be understood from a simple experiment. Consider a box containing many color balls from which we are drawing balls. If no single color predominates in the box, then our uncertainty about the color of the ball is maximal and the entropy is high. On the other hand, if the box contains black colored balls more than other colors, then there is more certainty about the color of a drawn ball, and the entropy is lower. Intuitively, a gambler would prefer the second case, because it is possible to place bets on black and win. In fact, in the extreme in which every ball is black, the entropy would be zero, and the gambler would win every time.

Entropy thus measures the average amount of information associated with a drawn ball. Intuitively, in the third case, the color of the ball is a certainty, and there is no information conveyed by knowing the color of a drawn ball. In the first case, knowing the color of previously drawn balls tells a gambler a lot of information about how she should bet.

Mathematically, given a discrete random variable X and it consists of several events x, which occurs with probability p_x, the entropy $H(X)$ is given by:

$$H(X) = -\sum_{x} p_x \log(p_x)$$

Looking at the above equation, there are two basic ways in which entropy can change:

(a) If the total number of events in X increases, entropy of X will increase. This is because entropy is defined as a summation of the values given by a function based on the probabilities of X. (Note that the negative log of a probability is always a positive number.)

(b) If distribution on X becomes more uniform, entropy will also increase.

Conditional Entropy

A more complex idea is the concept of conditional entropy. The conditional entropy *H(Y|X)* measures how much entropy a random variable *Y* has remaining if we have already learned completely the value of a second random variable *X*.

An easier way to understand conditional entropy is to first understand joint entropy, which is how much entropy is contained in a joint system of two random variables:

$$H(X,Y) = -\sum_{x,y} p_{x,y} \log(p_{x,y})$$

The conditional entropy can then be understood as:

$$H(X|Y) = H(X,Y) - H(Y).$$

METHOD

Viewing Bookmarks as Tuples

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* the set of users, and *T* the set of 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.

Data Collection

We collected del.icio.us bookmarking data using a custom web crawler and screen scraper. Our crawling tool simply 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. Given this, our methods assume, that the graph is fully connected, and to the extent that it is, our data is complete. However, since this walk was random, and limited in the amount of time we allocated to data collection, our data are most likely not a complete set of the del.icio.us data. We are however confident that our collection is a fairly-complete subset of the del.icio.us data. Table 1 shows the total number of distinct elements in our database.

With each bookmark tuple, del.icio.us stores the date on which it was bookmarked. This data gives us a means to "roll back the clock" and analyze the history and trends of bookmarking in over time.

DOCUMENTS	*USERS*	*TAGS*
9,853,345	140,182	118,456

Table 1. Total number of distinct elements in our del.icio.us datatbase.

Joshua Schachter started del.icio.us social tagging site in late 2003. The small amount of data we have for late 2003, however, appears to be somewhat noisy, so we excluded it from our analysis.

Figure 1 depicts the rate of growth in documents, users, and tags in the del.icio.us system. To our knowledge, this is the first published graph depicting the growth of the system.

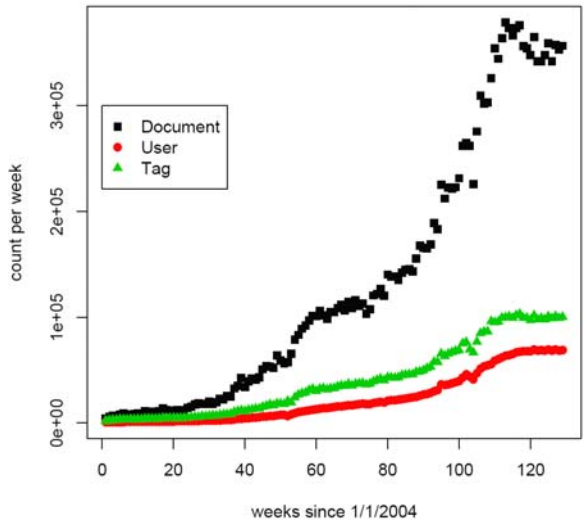

Figure 1. Graph depicting the rate of growth in documents, users, and tags over time. Note that the graph is plotted by count per week.

ANALYSIS AND RESULTS

Tags as a Vocabulary Describing Documents

We analyzed documents bookmarked in del.icio.us over time. 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 entropy the curve for the document set.

As shown in Figure 2, one can see that the entropy of the document set, *H(D)*, continued to increase. We know from data presented in the last section that the number of documents in the system is increasing, contributing to this increase in entropy. This means that, over time, users continue to introduce a wide variety of new documents into the system and that the diversity of documents is increasing over time.

A intuitive understanding of entropy of document *H(D)* increasing over time is that, if users were simply browsing the document set without any navigation aids, one would need a longer and longer identifier to navigate to a specific document.

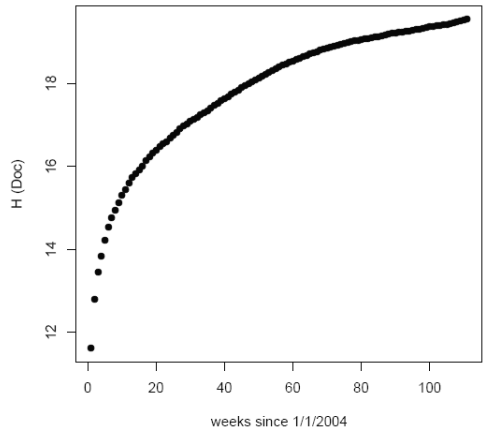

Figure 2. Entropy of documents H(D) is increasing over time.

Given our analysis of the entropy of the document collection, it only seems natural to do the same with the tag collection. Figure 3 shows a marked increase in the entropy of the tag distribution *H(T)* up until week 75 (mid-2005) at which point the entropy measure hits a plateau. At the same time, the total number of tags is increasing (see Figure 1), even during the plateau section of Figure 3.

Since the total number of tags keeps increasing, tag entropy can only stay constant in the plateau by having the tag probability distribution become less uniform. What this suggests is that users are having a hard time coming up with "unique" tags. That is to say, a user is more likely to add a tag to del.icio.us that is already popular in the system, than to add a tag that is relatively obscure.

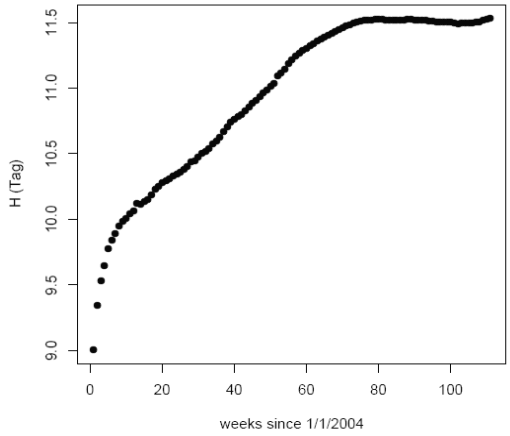

Figure 3. Entropy of tags *H(T)* is increasing at first, then started to plateau around Week 75 (mid-2005).

What's perhaps the most telling data of all is the entropy of documents conditional on tags, *H(D|T)*, which is increasing rapidly (see Figure 4). What this means is that, even after knowing completely the value of tags, the entropy of the document is still increasing. Conditional Entropy asks the question: "Given that I know a set of tags, how much uncertainty regarding the document set that I was referencing with those tags remains?" 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. That is to say, as a navigation aid, tags are becoming harder and harder to use. We are moving closer and closer to the proverbial "needle in a haystack" where any single tag references too many documents to be considered useful.

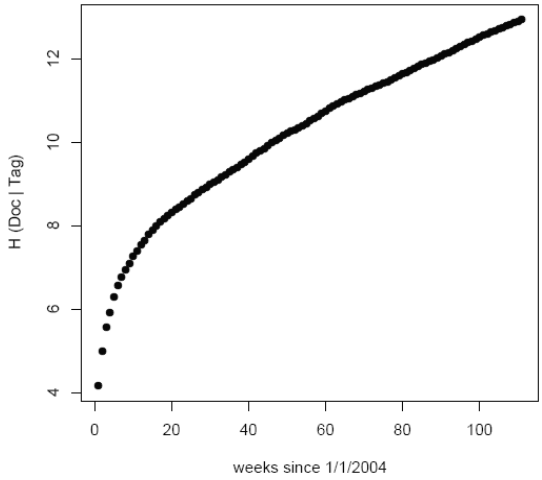

Figure 4. Entropy of Documents conditional on Tags *H(D|T)* increases over time.

More interestingly, we found that plotted on a log-log plot, the entropy forms a line (see Figure 5). Known as a power law, we can predict the growth of the conditional entropy *H(D|T)* for the next two years.

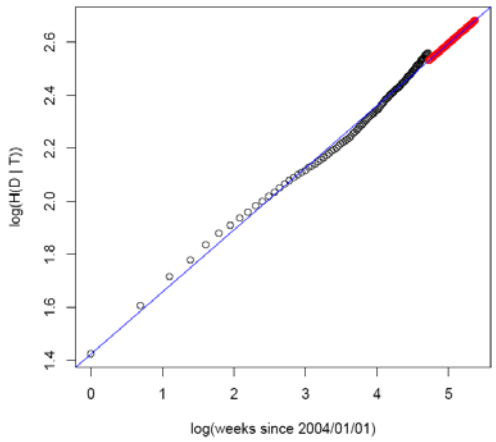

Figure 5. Conditional entropy *H(D|T)* over time forms a power law that can be predicted. Black shows current data, while the points in red show our prediction of conditional entropy.

What these results suggest is that even with a tagging system, the navigability of the document set is becoming more challenging overtime. One way for users to respond to this evolutionary pressure is to increase the number of tags they use to specify a document. Figure 6 shows the

number of tags per bookmark over time. The trend is clearly increasing, complementing the increase in navigation difficulty.

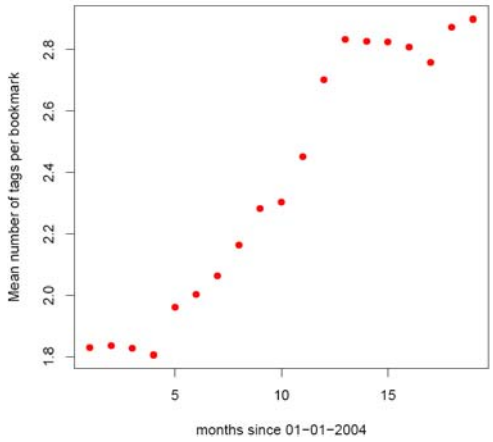

Figure 6. Raise in the number of tags per bookmark over time.

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 7.

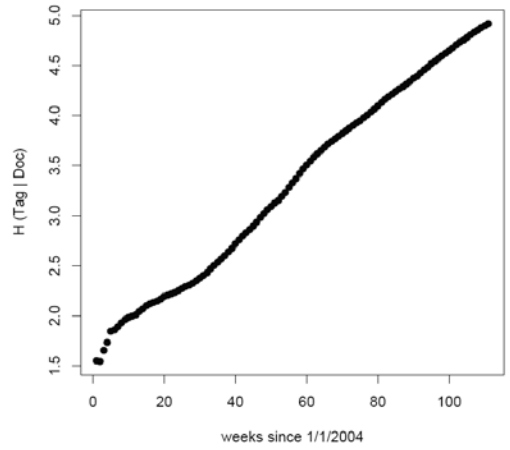

Figure 7. Conditional entropy *H(T|D)* increases over time.

Given a set of documents, we are seeing more uncertainty in the set of tags that is needed to describe those documents. This uncertainty results from either (1) number of tags is increasing or (2) tag distribution becoming more uniform, or (3) both. Either way, this means users are increasingly having a harder time specifying the tags for documents.

Relationship between User and Documents

So far, we have only used entropy to understand the trend of the relationship between tags and documents. Since we modeled a bookmark as a 3-tuple, we can also use our entropy methodology to understand the relationship between user and documents as well.

Figure 8 shows the entropy measure applied to the distribution of users. Relating back to the gambler, her bet would be placed on which user is going to make the next bookmark. It is easy to see that, much like the *H(D)* curve in Figure 2, the amount of uncertainty in which user is going to contribute a bookmark next is increasing.

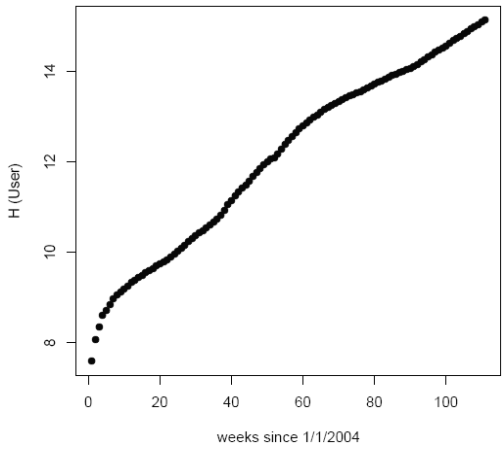

Figure 8. Entropy of user *H(U)* is increasing over time.

However, that is only part of the story, as one needs to look at the conditional entropy in order to gain insight into *what those users are tagging*. We really would like to ask the question "Given that I know a user U, how uncertain am I in the document that this specific user tagged"? Conditional entropy, *H(D|U)*, does just that.

The entropy of documents conditional on the user specifies how the diversity of documents have changed over time given you know a set of users. The results (Figure 9) show that the diversity of the documents that del.icio.us users have been bookmarking has plateaued. Given that we know the total number of users are increasing, and that the total number of documents are increasing, this data implies that users are increasingly more likely to bookmark the same content. In fact, this result shows that there is a large overlap in the del.icio.us bookmarks.

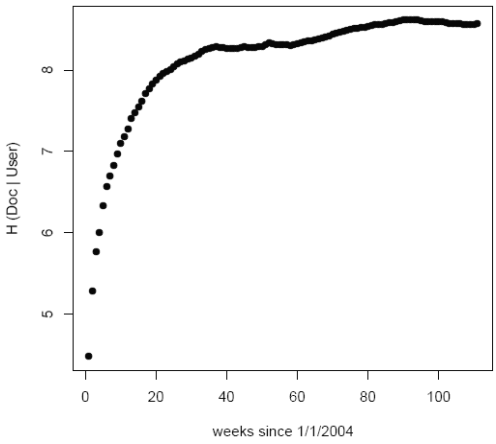

Figure 9. Conditional Entropy *H(D|U)* increases and then plateaus.

The entropy of users conditional on the documents *H(U|D)* specifies how diverse of a user set one would find if one knows the values of the documents. This entropy specifies how difficult it would be to find experts on a set of documents. Figure 10 shows that it linearly increases.

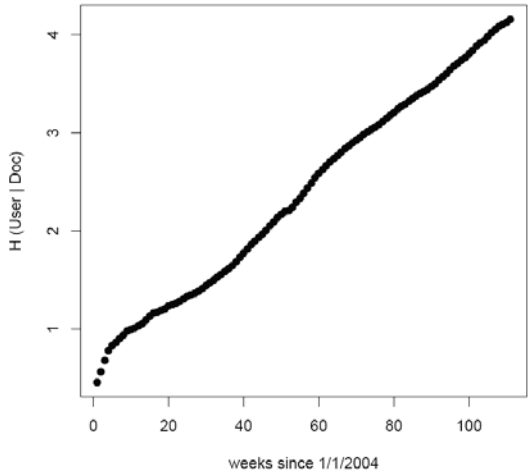

Figure 10. Conditional Entropy *H(U|D) increases linearly.*

Relationship between User and Tags

We can also look at the conditional entropy, *H(U|T)*, which specifies the amount of uncertainty in finding a set of users who uses a set of tags. In a way, this measure is also looking at the expertise problem: Given I have a list of tags, which users are the experts on these topics? As shown in Figure 11, it is harder and harder to find which users are the experts, given a set of tags.

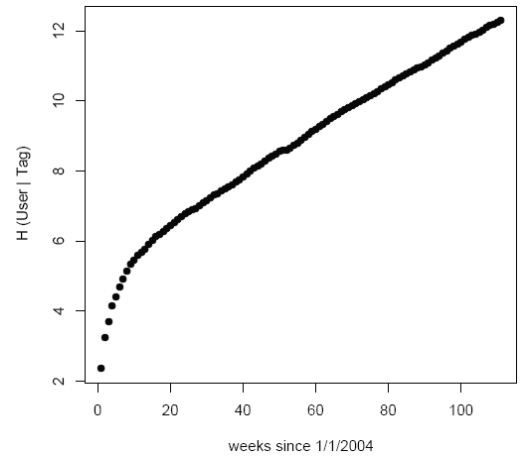

Figure 11. Conditional entropy *H(U|T)* increases rapidly.

Conditional entropy *H(T|U)* is the amount of entropy in tags given we already know the set of users. In a way, this is a user modeling problem. It specifies, given a set of users, how diverse of a tag set (topics) do we have? From Figure 12, we see that it increases gradually after about Week 20. This means it is getting slightly harder for a user modeling system to identify a set of topics that a set of users is interested in.

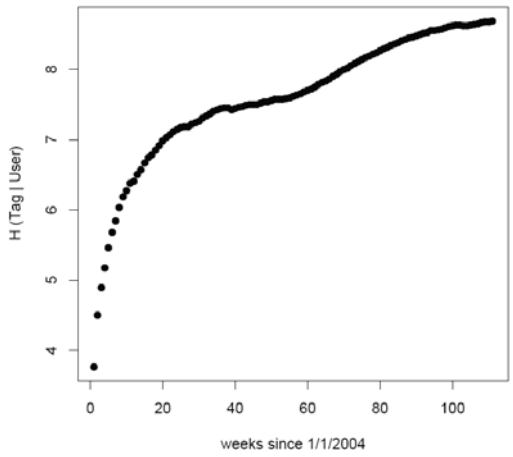

Figure 12. Conditional entropy *H(T|U)* increases gradually after Week 20.

Summary

The conditional entropies we examined can be used to understand the trends of a social tagging system from multiple perspectives:

- *H(D|T)* specifies the social navigation efficiency. How efficient is it for us to specify a set of tags to find a set of specific documents? We found that in del.icio.us that it is getting less and less efficient.
- *H(T|D)* specifies the social tagging efficiency. How efficient is a set of tags to describe a set of documents? In our context, how efficient is it for taggers to specify a set of tags to describe a set of documents? In del.icio.us, social tagging is getting less efficient.
- *H(D|U)* specifies the document recommendation efficiency. How easy is it for us to specify a set of documents for a set of users? We found that this measure hit a plateau, and that it is getting neither harder nor easier to recommend documents to users. This is also a measure of the degrees of overlaps of documents visited between users.
- *H(T|U)* specifies the user modeling efficiency. How easy is it to identify a set of tags (which specifies the topics) given a set of users? We found that it is getting slightly harder to find topic experts in del.icio.us, perhaps because more information diffusion is happening. Users are using a more diverse set of tags over time, which was independently verified by Golder and Huberman [10].
- *H(U|D)* and *H(U|T)* specifies the topic and document expertise efficiency. How easy is it for us to find a group of users, given we know either a set of documents or tags? We found finding user expertise is becoming more and more difficult.

DISCUSSION

According to this analysis, the set of social tags (which is a set of uncontrolled vocabulary generated by a crowd) is becoming more diverse over time, and this set of tags and their mapping to documents are slowly losing their ability to direct users to any specific document. That is, the social navigation afforded by the social tags in the system appears to be decreasing in their descriptive effectiveness.

Also, the collective of users on del.icio.us is increasingly having a harder time in tagging documents in del.icio.us. They are less certain what tags should be used to describe documents. A piece of evidence that is consistent with this observation is that users appear to have responded to this evolutionary pressure by increasing the average number of tags per document over time.

Vocabulary Problem Revisited

Coming back to the vocabulary problem, the issue from a social tagging system perspective can be broken up into two sides of the same coin. On one side, the vocabulary provides a way to describe documents (the social tagging problem). On the other side, the vocabulary provides a way to navigate through the documents (the social navigation problem).

We can see from our analysis that the vocabulary that is emerging in del.icio.us is becoming less efficient. This is somewhat understandable, since the amount of information being introduced into the system is growing at an extremely fast pace according to a power law. As the tags that people use to describe these documents become more saturated, they need to either find new words or they need to use more words to describe the contents. Indeed, overall we are seeing more and more new tags being introduced into the system. However, since only so many words are applicable to describing the content of a document, users cope by increasing the average number of tags they use. This is how the social tagging problem is playing out in del.icio.us.

Real usage data to analyze the social navigation problem on del.icio.us is unavailable to us, so it is hard to know how users are really browsing and searching through the tags, and how efficient they are. Our analysis above show that social navigation is becoming more difficult, and it would be good to verify this via comparison with navigation data from user logs.

However, interestingly enough, social navigation happens on search engines as well, and the data there correspond to our analysis. As more and more web pages get created and indexed, search engine users have increased the average number of search keywords used in search engine queries [1]. Yahoo!'s latest data shows that the average search query length has increased from 1.2 words in 1998, 2.5 words in 2004, to 3.3 words in May 2006. So web search engine users, responding to the evolutionary pressure of having more documents to wade through in search engines, have increased the specificity of their queries by increasing number of query words.

DESIGN IMPLICATIONS

Our methodology provides two main design implications:

First, from a system design point of view, this situation is clearly less than ideal, and we would like to increase the efficiency of the tags. Fortunately, these entropy measures provide a way to analyze the efficiency of an existing social tagging/navigation system, thereby providing a method for evaluating system design choices.

Second, the data suggest that the interest overlap of users is large, both in tagging vocabulary as well as in the actual documents they are viewing. There might be a way to take advantage of this overlap and create a search engine that enables more efficient social navigation.

Evaluation of System Design

One hypothesis of what makes social tagging so powerful appears to stem from the fact that the entire process is done in a distributed fashion, with each person bringing his/her own personal tag bias to the table while conforming to community influences [15]. The personal bias is one way in which a folksonomy can circumnavigate the vocabulary problem of standard organization systems. However, as Sen et al. noticed, over time, personal bias is influenced by popular tags in the system, thereby reducing the power of a personal vocabulary [15]. Indeed, there is a tension between keeping any single user's vocabulary personal, and having every user only use words that she feel are useful to the community at large. If a user only uses words that are meaningful to him/her then their contribution to the vocabulary at large is minimized. What is needed appears to be some ways to measure these personal biases and community influences.

Our methodology provides a way to gainfully measure these opposing forces, i.e. personal bias and community usefulness, of a vocabulary. Personal bias toward an individual vocabulary can be measured via $H(Tag|User)$. This is what we called the "user modeling efficiency". At the same time, collective usefulness of a vocabulary is defined by $H(Tag|Docs)$. This is what we called the "social tagging efficiency". These metrics provide a means for evaluation of design changes to a system like del.icio.us.

Yahoo! is interested in producing a collaborative tag suggestion engine. Indeed, removing the cognitive barrier to tagging is easily seen as a useful tool, especially for a company that now owns many of today's prevalent social tagging systems. In a workshop last year focused on collaborative tagging, Zhichen et al. from Yahoo! Research introduced a tag suggestion engine [21]. Their system has a few smartly chosen heuristics to evaluate the "usefulness" of a specific tag, and pre-computes a set of possible, suggested tags for every object in their system. A task left

for the researcher would be to apply our methodology to the vocabulary/document collection that this system would produce. The resulting entropy measures would give insight into how efficient the tag collection is in describing the document set.

Indeed, design changes and tools for social tagging are being suggested all the time. For example, on August 30th, 2005, John Resig introduced "a del.icio.us bookmarklet that auto-tags and auto-describes your bookmarks", appropriately named Lazy Sheep [19]. This tool makes sense on an individual level as it removes some of the cognitive barriers to tagging, however, from a community usefulness point of view, this tool, if used by a large number of bookarkers, has potential to remove individual vocabulary bias. If so inclined, del.icio.us could apply our methodology to analyze the impact of such a tool.

In a companion paper, we are exploring ways in which users could easily click on words in a paragraph and tag those paragraphs for later retrieval [12]. The idea is to enable within-page annotations. This tool would clearly generate much more tagging data, and we do not yet know how it would perform with many users using the tool. The methodology here could be applied to understand the efficiency of the system.

User-Item Overlap and Social Search

The second implication of our analysis shows that a large number of users are in fact overlapping in the documents that they bookmark. We have been interested in applying these findings and implications to the design of a social search engine. Indeed, the most popular documents on del.icio.us have been bookmarked by thousands of users.

With enough overlap in their tags and documents, users might be able to collaborate in their search process. As the social navigation efficiency analysis above suggests, there might be a surprising amount of overlap in bookmark data. We are performing some analysis of real user browsing histories and their bookmarks, trying to understand to what extent their information sources are diverse but yet overlap. The preliminary results show that, given the vastness of the web, it turns out that there is a lot more overlap than we might otherwise have expected. We are currently constructing social search technology based on these ideas.

CONCLUSION

Researchers are interested in understanding, characterizing, and evaluating the efficiency of social tagging systems [8], especially since social tagging systems have become popular. Given this popularity, we seek to understand the efficiency of tagging vocabulary emerging from these distributed social taggers on the web.

We analyzed a popular social tagging site, del.icio.us, using information theory. By analyzing various kinds of entropy, which is a traditional information theory metric, we found that, over time, del.icio.us is becoming harder to navigate. Moreover, the collective of users (the crowd) is having a harder time in tagging documents as the collection of bookmarks grows unabated. This is somewhat intuitive, since the amount of information being bookmarked is growing extremely fast, and the usage and growth of the tagging vocabulary become much more saturated. Entropy, as a metric, can also be used to drive system design choices. We discussed several social tagging tools or modifications to social tagging that could benefit from using entropy to evaluate the effects. It is our hope that HCI researchers will utilize this methodology to characterize future social and collaborative information systems.

ACKNOWLEDGMENTS

We would like to thank Peter Pirolli for some particularly enlightening conversation in which we checked over the basic approach used in this paper. We also would like to acknowledge the comments of members of the User Interface Research Group at PARC.

REFERENCES

1. Bogatin, D. Yahoo: ‘Searches more sophisticated and specific’. ZDNet Micro-markets Blog. May 18, 2006. http://blogs.zdnet.com/micro-markets/index.php?p=27. (Retrieved Sept 29, 2006).
2. V. Bouthors and O. Dedieu. Pharos, a collaborative infrastructure for web knowledge sharing. In ECDL, pages 215–233, London, UK, 1999. Springer-Verlag.
3. Burt, R.S. (2004) Structural holes and good ideas. American Journal of Sociology, 110(2): 349-399
4. CiteULike. http://citeULike.org. (Retrieved Sept 29, 2006).
5. del.icio.us. http://del.icio.us. (Retrieved Sept 29, 2006).
6. Flickr. http://www.flickr.com. (Retrieved Sept 29, 2006).
7. Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. The vocabulary problem in human-system communication. Commun. ACM 30, 11 (1987), 964-971.
8. 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 CHI '06 Extended Abstracts on Human Factors in Computing Systems (Montréal, Québec, Canada, April 22 - 27, 2006). CHI '06. ACM Press, New York, NY, 36-39.
9. N. Glance, D. Arregui, and M. Dardenne. Knowledge pump: Supporting the flow and use of knowledge. In Information Technology for Knowledge Management. Springer-Verlag, 1998.
10. Golder, Scott and Bernardo A. Huberman. (2006).

"Usage Patterns of Collaborative Tagging Systems." Journal of Information Science, 32(2). 198-208.

11. Hammond, T., T. Hannay, B. Lund, and J. Scott. Social bookmarking tools : A general review. D-Lib Magazine, 11(4), April 2005.
12. Hong L. and E. H. Chi (2006). ColTag: Supporting Social Annotations of Web Pages. (Under review by CHI'07 Notes.)
13. MacGregor, G. and E. McCulloch. Collaborative tagging as a knowledge organisation and resource discovery tool. Library View (accepted), 55(5), 2006.
14. D. R. Millen, J. Feinberg, and B. Kerr (2006). Dogear: Social Bookmarking in the Enterprise. *Proc. CHI'2006*, 111-120.
15. Sen, Shilad, Shyong K. Lam, Dan Cosley, Al Mamunur Rashid, Dan Frankowski, Franklin Harper, Jeremy Osterhouse, John Riedl. tagging, community, vocabulary, evolution. To appear in Proceedings of CSCW 2006.
16. Shirky, Clay. Ontology is Overrated: Categories, Links, and Tags. Blog entry. http://shirky.com/writings/ontology overrated.html (retrieved Sept 21, 2006).
17. Sinha, Rashmi. A cognitive analysis of tagging. September 27, 2005.
18. Vanderwal, T. (2005). Off the Top: Folksonomy Entries. http://www.vanderwal.net/random/category. php?cat=153 (Retrieved November 5, 2005.)
19. Resig, John. Lazy Sheep Bookmarklet. http://ejohn.org/projects/sheep/, August 30th, 2005. (Retreived Sept 28, 2006).
20. YouTube. http://www.youtube.com. (Retrieved Sept 29, 2006).
21. Xu, Zhichen, Fu Yun, Mao Jianchang, and Su Difu. Towards the Semantic Web: Collaborative Tag Suggestions. In Proc. of workshop on collaborative Web tagging. Edinburgh, Scotland. May 2006. http://www.rawsugar.com/www2006/13.pdf

.

Contribution Statement:

Develops a methodology for characterizing efficiency of social tagging systems using information theory. Can help designers of collaborative tagging systems to evaluate effects of design choices.