

Understanding the Efficiency of Social Tagging Systems using Information Theory

Ed H. Chi
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA
echi@parc.com

Todd Mytkowicz*
Dept. of Computer Science
University of Colorado, Boulder, CO
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 underlying document objects? Does this distributed process really provide a way to circumnavigate the traditional “vocabulary problem” with ontology? 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. We show that information theory provides a natural and interesting way to understand this efficiency—or the descriptive, encoding power of tags. Our results indicate the efficiency of tags appears to be waning. We discuss the implications of our findings and provide insight into how our methods can be used to design more usable social tagging software.

Categories and Subject Descriptors

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

Keywords

Social tagging, navigation, information access, ontology, efficiency, evaluation, methodology, information theory, entropy.

*Research done while a summer intern at PARC

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT'08, June 19-21, 2008, Pittsburgh, Pennsylvania, USA.
Copyright 2008 ACM 978-1-59593-985-2/08/06 ...\$5.00.

1. INTRODUCTION

For most people, tagging is a method of personal organization; tags are pulled from a personal vocabulary to describe personal objects. The process of tagging, then, consists of *encoding* objects with keywords that individuals find interesting so as to afford *retrieving* those very same objects at a later date with known keywords. Presumably, the vocabulary, or keywords, an individual uses to encode objects will later help with retrieval. However, users do not tag in a vacuum; in almost any social tagging system of sufficient size, there is a large overlap in both what objects people find interesting and what vocabulary the collective users employ to describe those objects. There has been recent research interest in understanding how the shared vocabulary generated by a social tagging system might be a better organizational system that is somehow more usable and navigable than traditional ontology and hierarchical organization of information sources [11]. We can try and understand a social tagging system then, as balancing two essential operations; building a social repository of the objects its users find interesting, and by those same users, organically evolving a shared vocabulary to describe those same objects.

Indeed, the accumulation of human knowledge relies on innovations in novel methods of organizing information. Subject indexes, ontology, library catalogs, Dewey decimal systems are just a few examples of how curators and users of information environments have attempted to organize knowledge. Surprisingly, the use of free-form labeling through tagging may appear like a recipe for disaster but instead it has turned into one of the newest trends on the web. Social tagging has been applied to photos (flickr.com), videos (youtube.com), web pages (del.icio.us), and academic paper citations (CiteULike.org), and in each of one of these cases, an organically-evolved, shared vocabulary, emerged to describe the content of the tagged objects.

But how does the newest fad in information organization compare to existing and established techniques? Clay Shirky argues that because tagging systems do not use a controlled vocabulary, they can easily respond to changes in the consensus of how objects should be classified [11]. Shirky's essay speaks to a potential advantage of encoding a diverse set of objects by a large number of users—both the repository and the vocabulary can evolve and change based on use. Furnas also noted this; he posits that generating a small set of keywords for an object is relatively easy for a single user, while generating a large set of keywords is a task best left for a group of users [3].

Furnas links social tagging to the vocabulary problem: “dif-

ferent users use different terms to describe the same thing”[4, 3]. MacGregor also refers to social tagging as, fundamentally, 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” [8]. Likewise, Sen et al. suggests: “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” [10].

All of this prior work alludes to a fundamental and unresolved issue in understanding social tagging systems: As researchers of information and hypertext systems, how do we study the efficiency of tags in organizing the objects they are supposed to encode? If tags *efficiently* encode navigational information for document objects, then users will more easily navigate to information sources in the system. If, on the other hand, tags are noisy, ambiguous, and often incorrectly applied, then users will have a hard time finding information in the system. This knowledge is paramount to understanding the operational use of tagging sites and at the same time might provide insights to designers of such systems. Indeed, to understand how tags have evolved for a large corpus of objects, such as in del.icio.us, we need to understand whether the tags adequately describe the objects 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 this paper, we suggest information theory as a natural framework in which to evaluate the efficiency of the vocabulary generated by social tagging sites. We provide information theoretic measures of both the encoding and retrieval process of tagging. We ground our work with an example analysis of a popular tagging site, del.icio.us.

This paper is organized as follows: Section 2 describes prior work. In Section 3 we detail our methods and data collection process. We also provide a short information theory primer. Section 4 demonstrates our information-theoretic analysis of del.icio.us. In Section 5 we generalize our work and discuss its implications while Section 6 concludes.

2. RELATED WORK

Social *bookmarking* systems have been a vibrant research area for quite a long time, including systems such as Glance et al.’s Knowledge Pump [5], Bouthors and Dedieu’s Pharos [1]. However, these systems did not use social tagging.

Tagging, or the use of free-form labeling to describe document objects, only recently became a focus of interest. Thomas Vander Wal described the vocabulary that organically grows from a diverse set of users as a “folksonomy” [12]. This term is now used to describe any Web-based technology for generating open-ended labels that categorize content collaboratively. The popularity of social tagging systems, perhaps, can be attributed to the benefits users perceive in the ease of encoding content and at a later date, recalling that very same content.

The surprising aspect of this phenomenon is that, in contrast to professionally developed taxonomies, a folksonomy appears rather unsystematic. And yet, as shown by other researchers, 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 [8]. Because many users can easily label content in a distributed fashion, many more objects are tagged. Moreover, because the process does not use a controlled vocabulary, it can easily respond to changes in the consensus of how users collectively think content should be classified. This is a point well-explored by Shirky in his essay [11].

Some academic research has focused on the dynamics of social tagging systems[6, 2, 7]. The most well-known is arguably Golder and Huberman’s work on understanding the usage patterns of social tagging systems [6]. Their work characterizes a small subset of del.icio.us and argues that there is a growth pattern to the tagging system. Moreover, they found that 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. MacGregor and McCulloch collected and discussed a set of arguments related to the pros and cons of controlled vocabulary vs. free-form labeling [8].

CSCW and HCI researchers have also started looking at the field as a area for investigation. Millen et al. introduced a system designed for tagging intranet items in an enterprise [9]. They collected some sample user data, and showed that users form social networks through patterns of their usage of others’ bookmarks. Sen et al. studied how tag selections are affected by community influences and personal tendencies [10]. They studied four different tag selection algorithms for displaying tags from other users and found that user tagging behaviors changed depending on the algorithm.

Within the HCI community, a CHI2006 conference panel organized by Furnas et al. brought social tagging systems to the attention of HCI practitioners [3]. Furnas’ suggestion that social tagging systems can be viewed from a “vocabulary problem” [4] perspective which 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. He indirectly suggested that 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. Taking his cue, we embarked on research that would use information theory to characterize this vocabulary problem in social tagging systems.

3. APPROACH

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, \dots, t_n . Without loss of generality, it is then possible to express the bookmark b as a set of 3-tuples $(d, u, t_1), \dots, (d, u, t_n)$. In our data, we decompose all of the bookmarks into this form. From this decomposition it is trivial to come up with the probability of a specific document ($P(D)$) or a specific tag ($P(T)$).

Table 1: Distinct elements in our del.icio.us database.

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

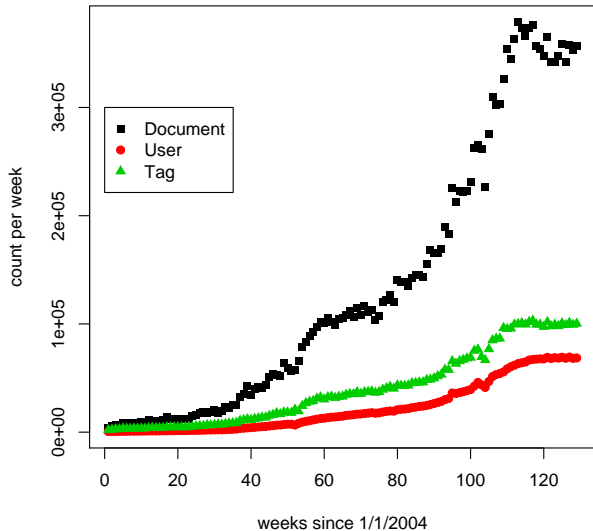


Figure 1: Graph depicting the rate of growth in documents, users and tags over time.

3.1 Data Collection

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 bookmarked the same document. In essence, our crawler did a random walk of the bi-partite graph of users and documents. This methodology mimics the process of crawling the web and much like crawling the web, our collection process could be biased and *may* under-represent the long tail of both users and documents (those users that infrequently bookmark as well as those documents that are infrequently bookmarked). Lacking the complete set of data for del.icio.us, however, we can only assume we have a representative and relatively complete sample. Our crawling process was done over a two month period, with some 40 machines. 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. **Figure 1** depicts the rate of growth in documents, users, and tags in the del.icio.us system.

3.2 Information Theory Background

For most tagging systems the total number of tags in the collective vocabulary is much less than the total number of objects being tagged. To put this in context, consider

our data from del.icio.us. The ratio of unique documents to unique tags is almost 84 (see Table 1). Given this multiplicity of tags to documents, a question remains: how effective are the tags at isolating any single document? Naively, if we specify a single tag in this system we would uniquely identify 84 documents— thus the answer to our question is "not very well!" However this method carries a faulty assumption; not every document is created equal. Some documents are more popular and important than others, and this importance is conveyed by the number bookmarks per document. Thus, we can reformulate the above question to be: how well does the mapping of tags to documents retain *information* about the distribution of these documents?

Information theory provides a natural framework to understand the amount of shared information between two random variables.

3.2.1 Entropy

Given a discrete random variable X which consists of events $\{x_1, \dots, x_N\}$, each of which occurs with probability $p(x)$, the entropy $H(X)$ is defined as:

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

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

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

The unit of entropy is the bit and is a useful metric for understanding the diversity of a random variable. To see why, consider that entropy is maximized when each event in the set X is equally likely, and minimized when one event takes on all of the probability mass. Any value in between gives a notion of the repeatability of the events being measured.

3.2.2 Conditional Entropy

Entropy is a measure, in bits, of the uncertainty in a single random variable. The conditional entropy, however, measures the amount of entropy *remaining* in one random variable when we know the value of a second random variable.

Given two discrete random variables $X = \{x_1, \dots, x_N\}$ and $Y = \{y_1, \dots, y_M\}$, such that the event $\{x, y\}$ occurs with the joint probability $p(x, y)$, the joint entropy:

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

Using this value, we can calculate the conditional entropy $H(Y|X)$, given by:

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

Conditional Entropy is a useful measure to help quantify the amount of shared information between two random variables. If X tells us *everything* we need to know to determine Y , the resulting conditional entropy $H(Y|X) = 0$. Conversely, if X tells us *nothing* about the random variable Y the resulting entropy will be $H(Y)$, and thus X and Y are independent.

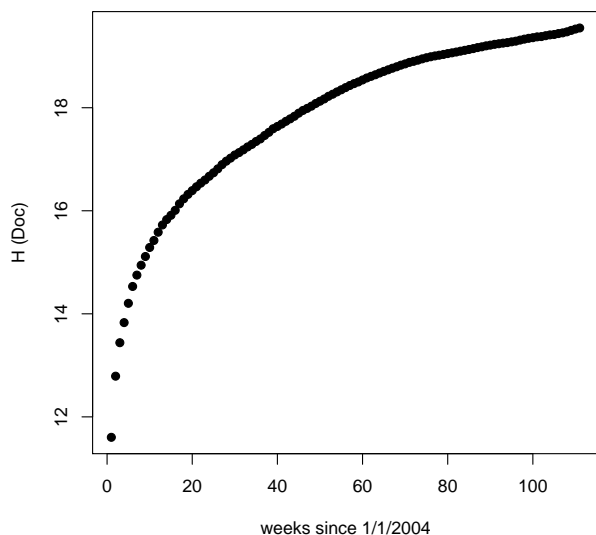


Figure 2: Entropy of documents $H(D)$ is increasing over time.

Because the definition of conditional entropy is not artificially bound to a preset range, conditional entropy can be a difficult measure to interpret. Consider $H(Y|X) = 10$; this result tells you very little about the independence between these two random variables, X and Y , unless you understand the amount of entropy in Y —conditional entropy is a relative measure. For this reason, practitioners usually use mutual information as a measure of independence.

3.2.3 Mutual Information

Given two discrete random variables $X = \{x_1, \dots, x_N\}$ and $Y = \{y_1, \dots, y_M\}$, such that the event $\{x, y\}$ occurs with the joint probability $p(x, y)$, the mutual information $I(X; Y)$ is given by:

$$I(X; Y) = H(Y) - H(Y|X)$$

Mutual information is a symmetric measure of the independence of two random variables. It is minimized ($I(X; Y) = 0$) when the two random variables are independent and maximized when two variables are completely dependent (e.g. $I(X; Y) = H(X, Y)$).

4. ANALYSIS OF TAGGING

The information theoretic measures described in Section 3.2 allow us to quantify, over time, (i) the diversity in tags and documents respectively and (ii) the amount of shared information between tags and documents. This analysis gives us insight into how effective tags are at encoding documents.

We first discuss the how the diversity of documents is changing over time.

4.1 Distribution of Documents

As shown in **Figure 2**, one can see that the entropy of the document set, $H(D)$, continues to increase. We know that the number of documents in the system is increasing,

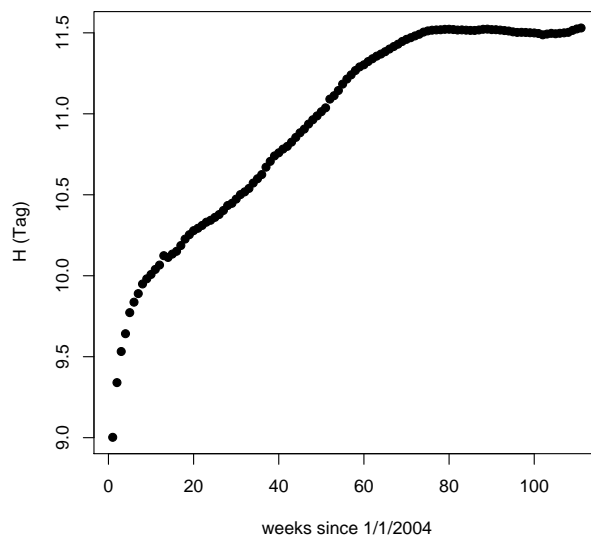


Figure 3: Entropy of tags $H(T)$ is increasing at first, then started to plateau around Week 75 (mid-2005).

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.

As the designer of a social tagging system, we would like to see more and more unique documents being introduced to the system, and indeed this is what is happening in del.icio.us. In essence, document growth has not stalled.

Given this fact, how are the usage patterns of the site changing over time? In the section that follows we answer this question. We break social tagging into two distinct processes and model them accordingly. In the first process, documents are *encoded* with tags by the collective set of users. Presumably encoding is done in such a fashion so as to afford the second process, *retrieval* at a later date.

4.2 Modeling Encoding

The amount of entropy that exists in the set of tags is a natural measure of diversity in those tags. Consider a social tagging site that only uses a single tag for each document: the diversity (entropy) of the tags is low. At the same time this fact eases the encoding process; all a user has to do is encode every document with the one, single tag, regardless of which document is being tagged.

Turning now to the data from del.icio.us, **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. During this same time period, the total number of tags is increasing (see Figure 1), even during the plateau section of Figure 2. Because 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 eventually the tagging vocabulary saturated, and coming up with new keywords became difficult. That is to say, a user is more

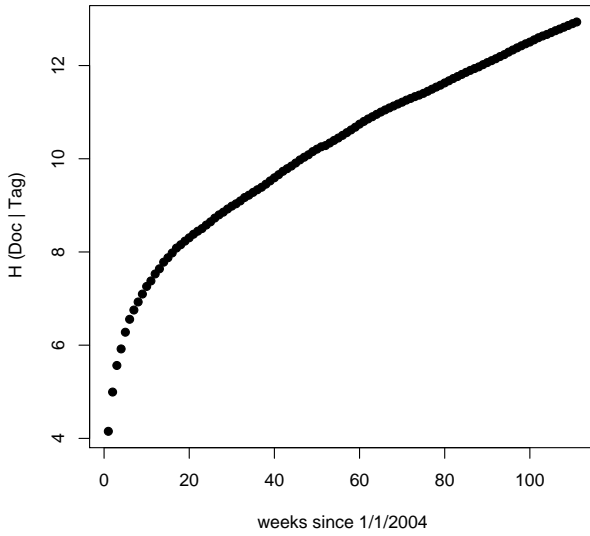


Figure 4: Entropy of Documents conditional on Tags $H(D|T)$ increases over time.

likely to encode an object with a tag that is already popular than to add a tag that is relatively obscure.

As the designer of social tagging systems, we expect users to employ a wide variety of tags to encode the document objects they are tagging. In the first 75 weeks of the system, this indeed is what was happening. After that, we saw less variety in the usage of tagging keywords. Given these results: (i) an ever increasing diversity in the document distribution and (ii) a saturating vocabulary, how is document retrieval affected?

4.3 Modeling Retrieval

The expected amount of entropy remaining in the document set *after* we know any individual tag is a good measure of the tag’s ability to accurately direct a user to any single document. Consider an example: every document in del.icio.us has a single, unique tag. In this situation, conditional entropy will be 0, as the tag, as a navigational aid, uniquely identifies a single document.

Given that we know the documents coming into del.icio.us are becoming more diverse, and at the same time the entropy for the vocabulary has stabilized, what effects does this have on being able to retrieve any given document with a specific tag? Put another way, does the encoding process accurately describe the content in del.icio.us and afford retrieval at a later date?

The entropy of documents conditional on tags, $H(D|T)$, is increasing rapidly (see **Figure 4**). What this means is that, even after knowing completely the value of a tag, the entropy of the set of documents is increasing over time. Conditional Entropy asks the question: “Given that I know a tag, how much uncertainty regarding the document that I was referenced with that tag remains?” 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,

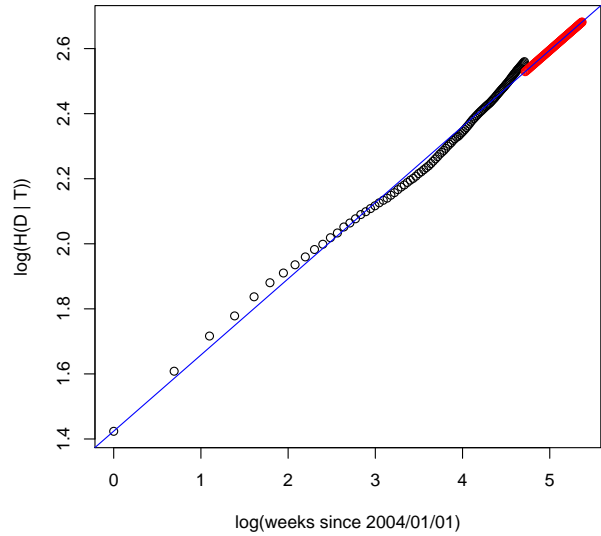


Figure 5: Conditional entropy $H(D|T)$ over time forms a power law that can be predicted. Black shows current data, while the points in red show our prediction of conditional entropy.

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.

One way to think about what $H(D|T)$ tell us is through the following scenario: A user asks the system to retrieve any document that has been tagged with the keyword “cooking”. When del.icio.us was young, back in 2004, it would return a small set of documents on the topic of cooking. Fast-forward to late 2006: if the user returns and repeats this retrieval process, using the same keyword “cooking”, she would find many more documents and the original set found back in 2004 would be lost in this mix. $H(D|T)$ gives us a measure for the amount of effort the user would have to expend to sort through the results in order to find any single document.

Interestingly, we found that when plot in log-linear scale, the conditional 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 few years. This phenomenon is most likely due to the regular influx of a diverse set of documents—or put another way—from the exponentially increasing size of del.icio.us over time. We note that this prediction is not absolute. For instance, if some external factor causes the dynamics of tagging to change this prediction will not hold. Section 5.2 illustrates one way in which the measures described in this paper can be used by architects of these systems to control the dynamics of both encoding and retrieval.

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

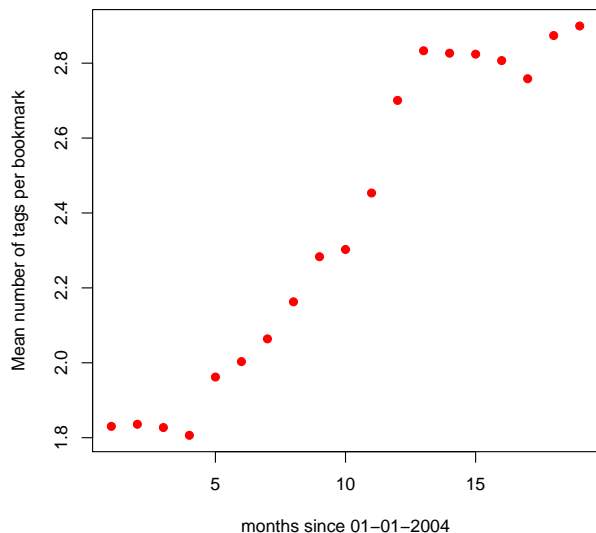


Figure 6: Increase in the number of tags per bookmark over time.

shows the number of tags per bookmark over time. The trend is clearly increasing, complementing the increase in navigation difficulty and the collective realization that a *single* tag is not a descriptive enough moniker to afford retrieval at a later date.

At week 100, (Figure 4) we see that the conditional entropy has reached a value of over 12 bits. Conditional entropy is a measure of independence (see Section 3.2.2), and we understand that an increasing $H(D|T)$ implies a loss of specificity in the encoding process of tagging. However, 12 bits is a rather hard number to interpret. To increase our understanding of this phenomenon, we investigated the mutual information of tags and documents.

To summarize our modeling of the retrieval process, we introduce mutual information. **Figure 7** illustrates the mutual information $I(D;T)$ of del.icio.us over time. As Section 3.2.3 mentions, mutual information is a measure of independence. Full independence is reached when $I(D;T) = 0$. As seen in Figure 7 the trend is steep and quickly decreasing. As a measure of usefulness of the tags and their encoding, this suggests a worsening trend in the ability of users to specify and find both tags and documents in the system.

5. DISCUSSION

This paper provides, to our knowledge, the first measure of the effectiveness of a vocabulary to encode a set of objects on a social tagging site. We have shown how information theoretic measures can be used to quantify the complex dynamics in a real, operating social tagging site.

Our analysis separates and distinctly models both the encoding and retrieval processes. When viewed as two different actions, it is interesting to note that these processes seem to be at odds with one another; any changes to the encoding

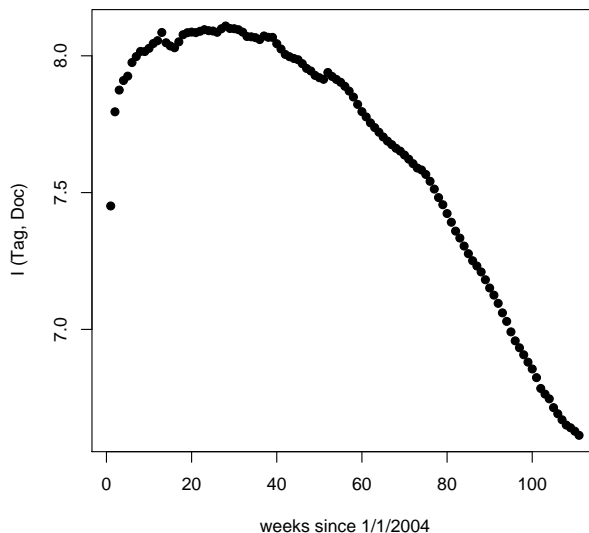


Figure 7: Mutual Information $I(T;D)$ decreases over time showing that tags are becoming less descriptive for any given document in del.icio.us. A value of 0 means complete independence between Tags and the Documents those tags are supposed to encode.

process will affect retrieval. Consider the following example where, for the sake of simplicity, we assume every document is just likely as every other document.

When $H(T)$ is minimized (only one tag is used for every document, for instance) the encoding process is very simple. However, this complicates retrieval—this single tag provides no information regarding any individual document as they are independent random variables. Thus, $H(D|T)$ is maximized. Conversely, if conditional entropy is minimal— $H(D|T)$ —each tag uniquely refers to an individual and specific document (each tag is a URL, for instance) and retrieval of any document is maximally efficient as documents and tags are completely dependent. Knowing a tag necessarily directs a user to a single document. This, however, implies that $H(T)$ is maximized because each tag is as likely as all others, and thus the distribution over tags is uniform. These results suggest that there exists an optimal balance between these two potentially competing processes. An open question remains: What is this balance and how does it change depending on the information needs of the individuals using any specific social tagging site.

5.1 Limitations and Future work

Originally, del.icio.us was not explicitly designed as a shared social tagging system. However, over time, it acquired features that encouraged social interaction. For instance, a shared vocabulary is developed organically through the use of tag suggestions for a particular document (eg. other people have tagged this with ...). Moreover, by exposing and encouraging browsing/searching of other users' personal repositories, a tagger can see what vocabulary other individuals use to tag their document collections. This also has

the effect of providing insight into what documents others find interesting. As we have shown in this paper, the outlook for shared tagging vocabulary systems seems troubling. Our results suggest that social tagging systems with shared vocabularies do not appear to enable users to navigate to any single document effectively. Fortunately, most of the time del.icio.us does not appear to be used as a shared social tagging site, because most users bookmark URLs for their own personal use. Admittedly, a limitation of our work here is that we did not analyze a shared bookmarking and tagging system like Ma.gnolia.com or Flickr.com. This is something we have left for future work.

We have made a simplifying assumption with our approach; our measures only consider the effectiveness of a *single* tag to retrieve a *single* document. Most documents, however, are tagged with multiple tags and thus may be uniquely identified by more than one tag. Indeed Figure 6 shows that the average number of tags per bookmark is around 2.8 as of late summer 2006. Future work includes investigating how this effects the conclusions we have drawn in this paper.

Another limitation here is that we did not analyze what portion of the entropy increases are due to increases in the number of tags and documents versus further unevenness in the tag and document distribution. Entropy as a measure of uncertainty is well-established in information theoretic research, but it would be nice to attribute what portions of uncertainty are caused by increases in the document set versus the diversity of the document set. One way to answer this question would have been to pick a subset of documents (or tags) in the beginning and see how the entropy for these documents (or tags) changes over time. We plan to do this analysis in the near future. Despite this our use of the entire set of documents in del.icio.us is not without merit.

A final issue to consider is that we essentially used a crawling method of link chasing. The samples acquired using this method may be somewhat biased toward our seed. Moreover, it is hard to obtain a complete tag network of del.icio.us or any other tagging site via these crawling techniques. Unfortunately, we have not been able to verify that we have a reasonable complete sample of del.icio.us at the time of our crawl, other than the fact that we were getting trickle amounts of data toward the end of our crawl. We hope engineers and researchers at del.icio.us could either provide valuable data sets to researchers like us for further study or start a research effort of their own.

5.2 Implications for Design

One question related to the implication of our work is how could these information theoretic measures be put to use so as to help designers understand what mechanisms to design in order to increase the effectiveness of the encoding process in social tagging systems?

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 users (we doubt any tagging site would want to do anything but increase this value!), all that can be controlled explicitly via the designers 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 designers of tagging sites are causing $H(T)$ to become less diverse. Instead, designers could provide an interface akin to the ESP

game [13]. 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. This may decrease the conditional entropy and thus provide a richer, more descriptive vocabulary for any given document.

When viewed in the light of information theory, social tagging can be is a form of mass dimensionality reduction; hundreds of thousands of diverse people are scouring the web for interesting documents and categorizing these documents with meaningful terms. The terms that people choose for the documents that catch their interests can be construed to capture the concepts that the document are about—the most salient aspects of its content. The lower dimensional representation people choose is an approximation of the full content of that very same document. Much like Latent Semantic Analysis (LSA) decomposes a set of documents it into a conceptual, lower dimensional latent space, so too does tagging. Given this observation, we posit that quantifying the amount of shared information between documents and tags, as we have done in this paper, is paramount to understanding the processes that are behind the dynamics of social tagging sites.

6. CONCLUSION

Much of tagging that happens on web sites are intended for the individual user. An individual employs a personal vocabulary to describe personal objects. However, organically through the efforts of many diverse users, a global language is developed that is used to describe the global set of objects. Other shared social tagging sites such as Flickr.com assume a global name-space that everyone can access.

In either case, as a large body of prior work shows, social tagging is fundamentally a method of organizing objects for later use. It is a process of *encoding* objects with keywords so as to later *retrieve* those very same documents. This retrieval could be done by the same person that encoded the object, or could be done by other users of the system. What prior work has not done is evaluate the effectiveness of this encoding process.

In this paper we suggest an information theoretic view of tagging. We provide a direct measure of encoding efficiency and apply this methodology to a popular social tagging site, del.icio.us. Our results show that with the gain in popularity of social tagging, the effectiveness of tags to refer to individual objects is waning. We conclude by detailing how our measures can directly used to help increase the effectiveness of social tagging sites.

7. ACKNOWLEDGMENTS

We would like to thank Peter Pirolli for some particularly enlightening conversations. We also would like to acknowledge the comments of members of the Augmented Social Cognition Research Group at PARC.

8. REFERENCES

- [1] V. Bouthors and O. Dedieu. A collaborative infrastructure for web knowledge sharing. *ECDI*, pages 215–233, 1999.
- [2] E. Chi and T. Mytkowicz. Understanding navigability of social tagging systems. In *SIGCHI alt.chi*, 2007.
- [3] G. W. Furnas, C. Fake, L. von Ahn, J. Schachter, S. Golder, K. Fox, M. Davis, C. Marlow, and

- M. Naaman. Why do tagging systems work? *CHI Extended Abstracts*, pages 36–39, 2006.
- [4] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 11:964–971, 1987.
- [5] N. Glance, D. Arregui, and M. Dardenne. Knowledge pump: Supporting the flow and use of knowledge. *Information Technology for Knowledge Management*, 1998.
- [6] S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 2006.
- [7] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 211–220, New York, NY, USA, 2007. ACM.
- [8] G. MacGregor and E. McCulloch. Collaborative tagging as a knowledge organization and resource discovery tool. *Library View*, 2006.
- [9] D. R. Millen, J. Feinberg, and B. Kerr. Social bookmarking in the enterprise. *CHI*, 2006.
- [10] S. Shilad, S. K. Lam, D. Cosley, A. M. Rashid, D. Frankowski, F. Harper, J. Osterhouse, and J. Riedl. tagging, community, vocabulary, evolution. *CSCW*, pages 181–190, 2006.
- [11] C. Shirky. Ontology is overrated: Categories, links and tags, September 2006.
- [12] T. Vanderwal. Off the top: Folksonomy entries, 2005.
- [13] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM.