Building a Better Folksonomy:
Web-based Aggregation of Metadata
CS6125 Paper

Blake Shaw <bs2018@columbia.edu>

October 10th, 2005

1 Abstract

We live in an age flooded with information. New technologies are making available many large unstructured sets of information. As this information becomes more available, it becomes more difficult to navigate without a guide. Now that a typical user can carry around 10,000 songs in his pocket, the choice of picking which song to listen to becomes increasingly more difficult. Now that a typical user can access 13 billion websites, how does a person know which sites are relevant to him?

The solution to this problem resides in building new web-based technologies that aid in the formation of folksonomies. Folksonomy is commonly defined as a large group of people spontaneously cooperating to organize information into categories [24]. Many websites today are taking advantage of the organizational powers of folksonomies, such as Wikipedia, Flickr, Technorati, Del.icio.us, Yahoo!, and others. All of these sites employ a simple tagging mechanism, where users attribute words or phrases to content. When these tags are aggregated, new metadata for that content is created.

Tagging offers amazing possibilities for information retrieval by using collective social intelligence to organize information instead of relying on one person's description or categorization. However, tagging only begins to approximate an ideal folksonomy. By simplifying the ways in which we collect metadata from the user, coupling this information collection more strongly with a social framework, and providing more powerful tools for categorization, we should be able to greatly improve systems for retrieving relevant information.

2 Introduction

Currently many groups are trying to solve the problem known as "information overload," first coined by Alvin Toffler. It is defined as having too much information to make a decision or remain informed about a topic [25]. In recent years, "tagging" has emerged as a new approach to the problem of finding relevant information on the web. Instead of mining the web for clues about which pages are relevant to whom, as search engines do, tagging relies on the idea of folksonomy. Many people categorize small parts of the web; when all of these tags are aggregated, a comprehensive set of metadata is produced which is directly related to how people think information on the web pertains to a certain topic.

2.1 Tagging: A Simple Folksonomy

2.1.1 The motivation for better metadata

Metadata is essential for effective categorization of information. Metadata is typically defined as data about data, information which classifies other specific information in more general terms. Traditionally, metadata is created by either the authors of the content or professionals who organize informational content as their job. Both of these traditional methods are inadequate in the sense that they require a single individual to know how the content is going to be used. As Mathes expands on this idea:

> Author created metadata may help with the scalability problems in comparison to professional metadata, but both approaches share a basic problem: the intended and unintended eventual users of the information are disconnected from the process. [13]

Furthermore, neither of these systems for producing metadata adapt well to a rapidly changing set of information such as the web. Tagging is one of the first mainstream solutions which tries to explicitly collect metadata from the end-users in a simple way.

2.1.2 How does tagging work?

The methodology behind tagging is simple: label content with keywords. For example, one might tag a vacation photo with terms such as "skiing," "mt. snow," "friends," etc. Tagging is introduced as a personal way to organize one's data, so later one can simply type "skiing" and pull up all of one's photos related to skiing. Although the act of tagging is very simple, the ways in which we can aggregate tags across users and content can be very complex. It is essential however, that the act of classification for the end-user is as simple as possible without compromising flexibility.

2.1.3 Aggregating tags creates a folksonomy

The term "folksonomy" was first coined by Thomas Vander Wal and is commonly considered "a neologism for a practice of collaborative categorization using freely chosen keywords" [24]. Figure 1 is a diagram that Vander Wal himself uses to describe this idea of a "Broad Folksonomy" as opposed to a narrow one, where tags are shared among users, and tagging is both a personal organizational scheme and a contribution to a general classification of some content.

We see from the diagram that a group of users, most of whom share a common vocabulary, classify an object with similar tags. This framework provides not only a searchable/browseable guide to content by means of entering keywords, but also creates connections between items, people, and tags in the following ways:

- Items that are tagged with similar tags can be considered similar and can be clustered together.
- People who use similar tags can be considered similar, which implies that content that is interesting to one person will be interesting to a similar person as well.
- Tags that are often used together can be considered to be related. In some cases they could be synonyms or hierarchically related by specificity.

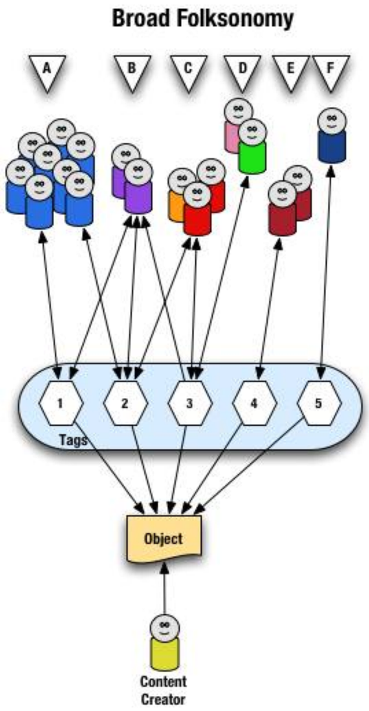


Figure 1: A diagram from Vander Wal's website explaining how tagging forms a broad folksonomy [22].

We will see in the following sections how current services take advantage of these 3 kinds of connections, and furthermore, how these connections can be better utilized.

2.2 Information Retrieval: Tagging vs. Search Engines

Search engines provide another method for creating metadata. By mining the inherent structure of the web itself, search engines are able to generate metadata about resources on the internet. In contrast to tagging where the end-users explicitly create metadata, metadata from search engines is created implicitly.

2.2.1 Search engines: extracting tags from links

The web contains in itself an implicit organizational structure created by how pages link to each other. Kleinberg eloquently poses the motivating question behind most search engines:

> The Web can be naturally modeled as a directed graph, consisting of a set of abstract nodes (the pages) joined by directional edges (the hyperlinks). Hyperlinks encode a considerable amount of latent information about the the underlying collection of pages; thus, the structure of this directed graph can provide us with significant insight into its content. Within this framework, we can search for signs of meaningful graph-theoretic structure; we can ask: What are the recurring patterns of linkage that occur across the Web as a whole? [11]

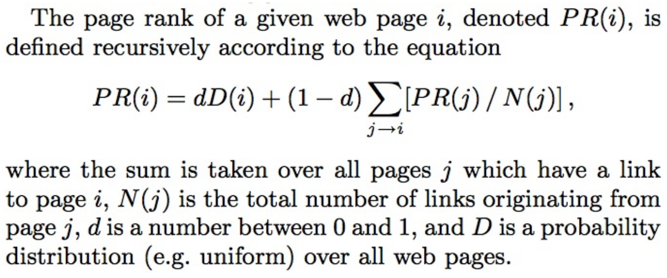
Search engines are in essence generating tags from links. For example, if on many different sites, www.apple.com is linked by the phrase "apple," "apple" becomes a tag which when searched reveals that website. The advantage to this system is that metadata is automatically produced. However, the disadvantage is evident when we consider where the metadata is really coming from: people who post links on websites. With the recent advent of blogging, this inadequacy becomes less of an issue: it is becoming easier for an average user to contribute links on the web. However, one can imagine a system which generates metadata from the large number of users who are surfing the web, not only the ones who are creating it.

2.2.2 Browsing vs. Searching

Traditionally, the purpose of a search engine is to enable users to find a specific thing which they are looking for, such as the website for a computer company, or information about a product. However, search engines are generally being used more and more to browse through information about a general topic, delving into more specific topics. In this case, "search" isn't the correct term, and search engines typically aren't suited to the task (although Google suggest beta is an interesting step towards this idea). A clear advantage to a tag-based system is that the tags can be thought of as filtering layers; one can add more layers incrementally to explore a topic. In terms of search, the analog would be presenting the user with the top rated other keywords the user could use to refine their search in different ways. More generally, tagging systems should be able to provide a better method for browsing by means of different kinds of information filters: people, websites, tags, etc. Combining sets of these items can filter an unfathomable amount of information into a relevant subset which can easily be browsed, personalizing the web into a smaller subset suited to an individual and certain topics.

2.3 Hubs and Authorities

In generating these tags, either automatically through a search engine, or collaboratively through a folksonomy, an important question emerges, how do we determine which sites are authorities on a given tag? Some notion of the reputability of a site is needed to sort search results by relevancy. Determining authorities and hubs from the link structure of the web alone is currently an interesting research topic. Google's PageRank algorithm is famous for its ability to provide better search results by incorporating a reputation metric. See figure 2.



The page rank of a given web page i, denoted $PR(i)$, is defined recursively according to the equation

$$PR(i) = dD(i) + (1-d)\sum_{j \to i}[PR(j)\,/\,N(j)]\,,$$

where the sum is taken over all pages j which have a link to page i, $N(j)$ is the total number of links originating from page j, d is a number between 0 and 1, and D is a probability distribution (e.g. uniform) over all web pages.

Figure 2: A simplified view of Google's reputation metric [4].

In essence the algorithm works by recursively giving each page a rank by seeing how many times it is linked to by other sites with a high PageRank. In more formal terms, PageRank can be considered the principal eigenvector of the adjacency matrix formed by the link structure of the web [11] [26].

This notion of identifying reputability is essential for collaborative tagging systems as well, although is only recently becoming implemented in current systems. Szekely proposes an idea of UserRank and TagRank, and he discusses implementation details for these algorithms [21]. More generally Russell nicely summarizes the goal behind such a system:

> Contextual Authority Tagging is the use of folksonomies to discover and define cognitive authority through reputation within communities of users. Authority is granted by individual users to other individual users with regard to their perceived domains of knowledge via free text tags or labels. This allows discovery of at least two things, 1) which users in a group are authority figures on a certain topic area, and 2) what areas of expertise a particular user possesses. [17]

Furthermore, CollaborativeRank is a relatively new site, which applies a UserRank algorithm to the Del.icio.us service [9]. For a more in depth look at Del.icio.us see the section comparing Del.icio.us and Yahoo! MyWeb 2.0.

2.3.1 The power law

The distribution of tags among users is an important characteristic of tag-based systems which needs to be utilized in order to properly identify hubs and authorities and perform other analysis. Similar to the link-structure of the web itself, the structure of links between tags forms a scale-free network that is subject to a power law distribution. Figure 3 and Figure 4 illustrate this phenomena.

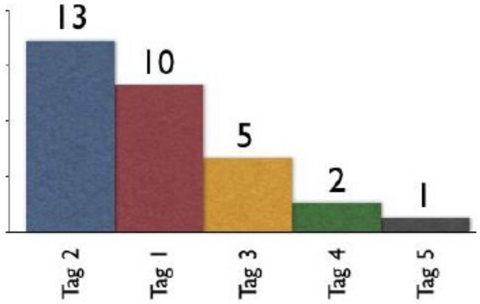


Figure 3: From the users tagging an object in Figure 1, we see a rough approximation to the power law distribution. [22].

Simply put, certain tags are used exponentially more than others, and certain users are much more important to certain tags than other users. These tags and users act as hubs in the networks. Links are not normally distributed because of the existence of these hubs that naturally emerge.

> Scale free networks have double-logarithmic diameter, i.e. an average distance between pairs of nodes grows as log log n. This non-trivial feature is explained as follows: while

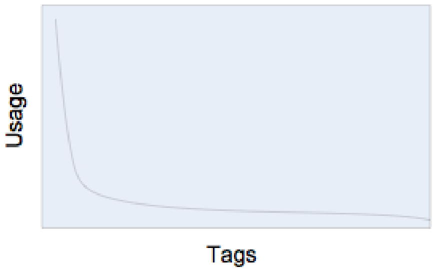


Figure 4: The power law distribution of tag usage [21].

> an average node has few links, hubs have a lot. So, a probability that your neighbor is a hub is high. Thus, hubs appear early in the shortest path spanning tree and fan out the tree, speeding up its exponential growth. [8]

The power law distribution has many interesting implications for an information retrieval system. Most important, it means that there is a small subset of keywords which are strongly and commonly agreed upon and can therefore reduce the shortest path to a specific piece of information dramatically.

3 Current Systems

Many groups are now taking advantage of a large group of people tagging information. This body of information does not need to simply be the WWW; however, the web is the primary focus of this paper. The musical catalog is another large source of information which needs to be categorized. Here is a brief survey of the different sites currently popular, including how their tagging methods are similar or different and how their system relates to folksonomy.

- **Flickr** is a service aimed to organize one's photo collection. They use tags as a personal organizational scheme; however, the service can be classified as a "narrow folksonomy" since not all users can tag content they see.

- **Wikipedia** is an online encylopedia. It is not a tag-based system. However, the service goes far beyond the typical convention of folksonomy. Where as in a typical folksonomy the goal is to simply organize information, the goal of Wikipedia is to create the information about a topic collectively from many users; the organizational scheme is produced by the links between topics.

- **AudioScrobbler** is a tool for studying what music people are listening to. It employs a tagging system where users can categorize songs, albums, artists with keywords for genre, or emotion, etc.

- **Del.icio.us**, **Technorati**, **Connotea**, and **Furl** are all systems for tagging content on the web through the idea of bookmarks. Each user creates a set of bookmarks, and tags them with descriptive keywords.

- **Yahoo! MyWeb 2.0** is a system for tagging bookmarks as well; however, the service also adds in more of a social framework. For more information see the following section discussing the service in more depth.

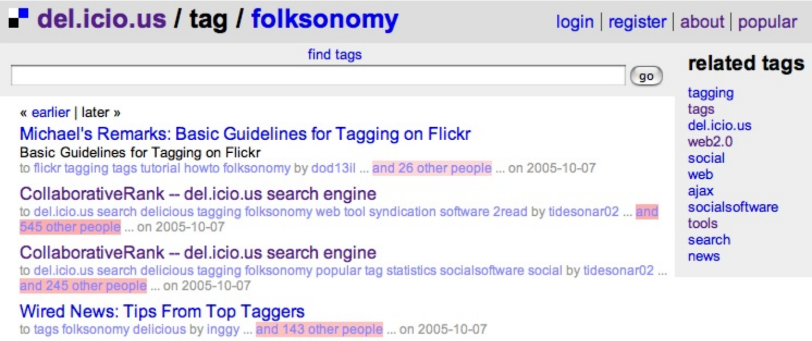
3.1 Two examples: Del.icio.us and Yahoo! MyWeb 2.0

To better understand current tagging implementations, here is a comparison of Del.icio.us and Yahoo! MyWeb 2.0. Both use bookmarks and simple words and phrases to create a folksonomy, with the goal of a set of collective bookmarks with smart, aggregated metadata.

3.2 Criteria and evaluation

Evaluating the usability of web-based services such as these is a difficult task due to the subjective nature of determining whether the system provides useful information to its users. Instead these services will be evaluated under the criteria of maturity, specifically how well they incorporate the idea of folksonomy in terms of the 3 kinds of information that can be utilized (see section 2.1.3).

3.3 Del.icio.us: simple bookmark tagging



Michael's Remarks: Basic Guidelines for Tagging on Flickr
Basic Guidelines for Tagging on Flickr
to flickr tagging tags tutorial howto folksonomy by dod13il ... and 26 other people ... on 2005-10-07

CollaborativeRank -- del.icio.us search engine
to del.icio.us search delicious tagging folksonomy web tool syndication software 2read by tidesonar02 ... and 545 other people ... on 2005-10-07

CollaborativeRank -- del.icio.us search engine
to del.icio.us search delicious tagging folksonomy popular tag statistics socialsoftware social by tidesonar02 ... and 245 other people ... on 2005-10-07

Wired News: Tips From Top Taggers
to tags folksonomy delicious by inggy ... and 143 other people ... on 2005-10-07

Figure 5: del.icio.us – A popular service for tagging bookmarks.

Del.icio.us was created by Joshua Schachter, who refers to the service as a "social bookmarks manager" [7]. The concept behind Del.icio.us is very simple, users surf the web, and when they see a site they want to remember, they click a Del.icio.us bookmark button, and type in a few tags describing the site. A user's bookmarks are then sorted by tags for easy retrieval. The remarkable aspect of this system is that it is entirely open. A user's bookmarks are publicly displayed at http://del.icio.us/*username*/, and all bookmarks for a tag are displayed at http://del.icio.us/tag/*thetag*/. Furthermore there is a page for the days most popular items, for both all bookmarks, and bookmarks relating to specific popular tags.

3.3.1 Analysis of interesting characteristics

There have been a number of studies of the Del.icio.us service which confirm predictions about the power law distribution and provide some other interesting analyses of other characteristics of tagging systems. Figure 6 shows that while certain power users have a large number of tags, there exist many users who use only a small set of tags. This “long tail” is characteristic of the power law distribution.

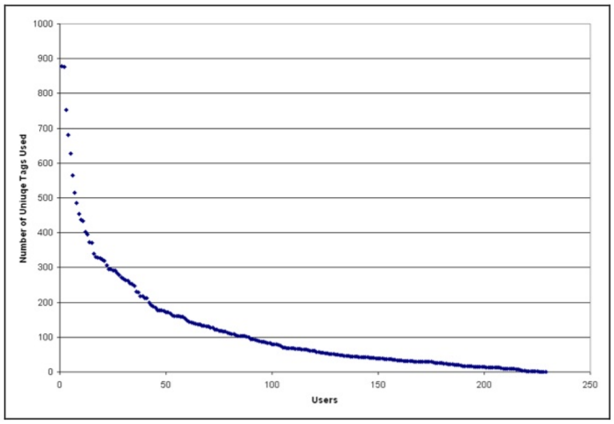


Figure 6: Number of tags per user. [7]

Furthermore, for each user there exists a power law distribution of the usage of their tags. As Shirky describes Figure 7:

> This is a single user’s tags. From here, you can tell something about this person – he or she is obviously a Flash programmer – the commonest tag here is Flash, followed by a number of other frequently used tags mainly related to programming. Like the front page, this distribution has the organic signature. Experts don’t catalog this way; experts who learn how to catalogue produce much more consistent labeling. Here, it’s whatever the user thought would help them remember the link later. [19]

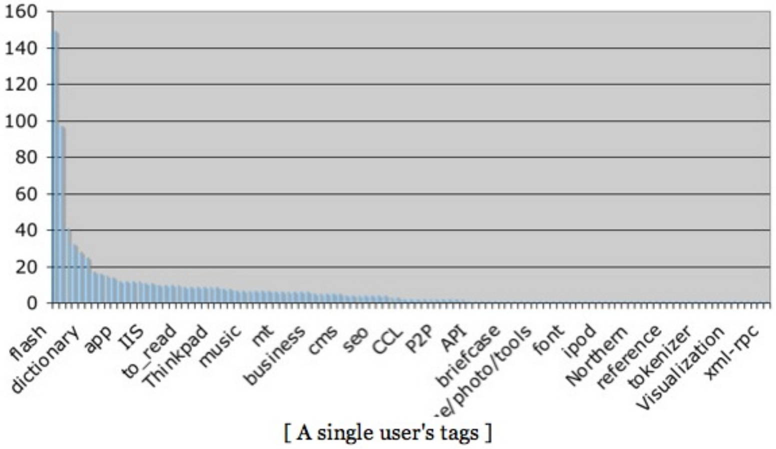


Figure 7: Frequency of tags for a given user. [19]

Another interesting characteristic of the Del.icio.us system is that tag proportions stabilize over time [7]. As we see from Figure 8, tags aggregated from many users stabilize to form stable proportions. This property is important in that it shows that only a small number of bookmarkings are needed to solidify a website in its proper place in terms of its tags, and also that these stable states may represent fundamental relationships between an object and the ideas represented by the tags.

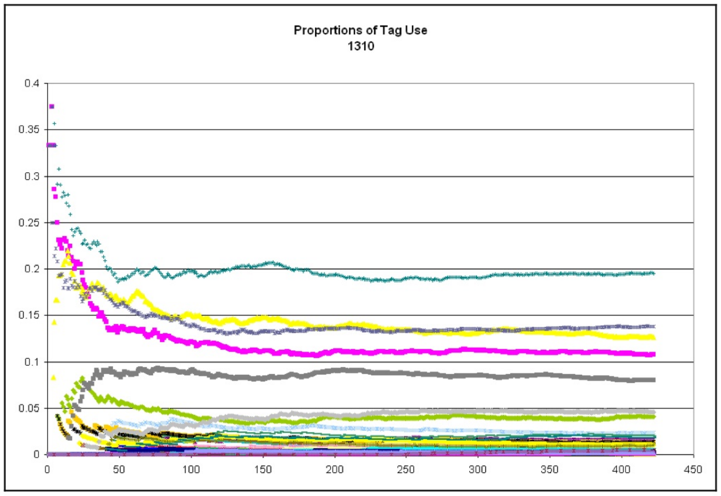


Figure 8: "The stabilization of tags relative proportions for popular URL (1310). The vertical axis denotes fractions and the horizontal axis time in units of bookmarks added." [7]

3.3.2 Conclusions

Del.icio.us has succeeded in creating a popular folksonomy for bookmarks and web content. The key to their success lies in the simple motivating factors for the users. They offer a simple way for a user to organize their bookmarks online, so that they can be easily searched and accessible from any computer on the internet. However, the aggregating features in Del.icio.us are very basic; they do not explicitly take advantage of the structural analysis that is possible for a tagging system. Other sites, such as CollaborativeRank, have introduced notions of hubs and authorities into the Del.icio.us system; however, Del.icio.us itself only provides basic functionality. In terms of the 3 types of information created by a tagging system (see section 2.1.3), Del.icio.us focuses only on the first: clustering similar items by tag.

3.4 Yahoo! MyWeb 2.0: Adding a social framework

Yahoo! MyWeb 2.0 beta is a newer service than Del.icio.us. It offers the same basic functionality as Del.icio.us: an easy way to keep track of one's bookmarks online, using tags. However, Yahoo! has included other interesting features, including a social framework, which takes better advantage of the properties of a tagging system and the information which can be generated by users tagging bookmarks.

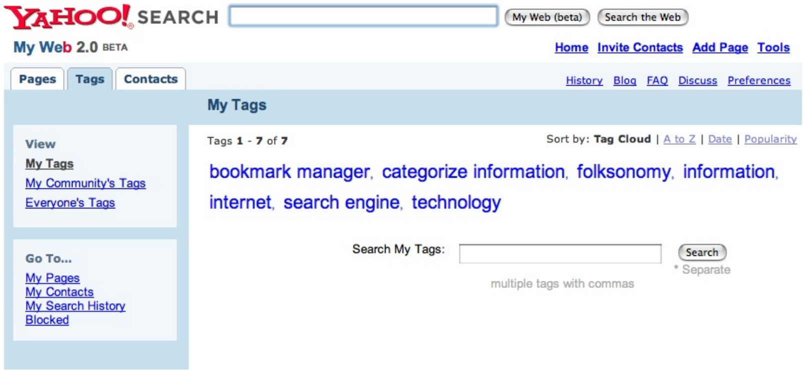


Figure 9: Yahoo! MyWeb 2.0 – A new bookmark tagging service with more of a social framework.

3.4.1 Tags as filters

Although similar functionality exists in del.icio.us, namely the ability to search for intersections of tags, Yahoo! explicitly builds into their interface a simple way to add/remove tags as away to filter information. They show a related tag list where one can add and remove selected tags in order to narrow the scope of the content one is looking for. In Del.icio.us, related tags are presented, but they are shown as a way to move from tag to tag, not to combine. One must explicitly search for tag intersections. This feature is significant since it is a step towards the goal of being able to filter content by tags, people, sites, etc. thus allowing the user to "weight" the web to better suit themselves.

3.4.2 MyRank: augmenting search with a profile

Another interesting feature of Yahoo! service is that they use the information about your tags and your community to help rank your web searches. In this sense, they are taking better advantage of the hubs and authorities identified by your tags, and your friends' tags, to help rank search results by relevancy to the user.

3.4.3 The social framework

Most importantly, Yahoo! incorporates the notions of friends and community into their tagging system. Users can invite other users via an email web form to enter their community. Community bookmarks provide an intermediary step between personal bookmarks and everyone's bookmarks. This addition is a vital step in personalizing content towards an individual. However, the system for adding friends in Yahoo! is too complicated since it relies on their Yahoo! 360 service. For the feature to be really useful, users should be able to simply add other users and tag them, much as they do with the content itself, as opposed to going through an email invitation process.

3.4.4 Conclusions

In terms of the criteria of maturity, Yahoo! incorporates a more complete notion of folksonomy in that the service takes advantage of both clustering items based on tags and showing people related to a certain tag (although not explicitly). Because of this additional social framework, Yahoo! is much closer to taking full advantage of the information produced by a tag-based system than a simpler system, such as Del.icio.us. However, neither system, truly takes advantage of the social nature of tag-based bookmarking system; neither system simply presents similar users, tags, and bookmarks for a given site, tag, or person.

4 Opportunities for Improvement

There are 3 main areas to improve these systems for aggregating user responses to provide more relevant information retrieval.

4.1 Make information collection more transparent and intuitive

Should we constrain ourselves to simply analyzing bookmarks in order to organize the web? Creating a bookmark and tagging it is somewhat labor-intensive. Are there more automatic ways to get useful information? What about a user's web-surfing history? What he has searched for? How he he found what he was looking for? All of these questions condense into a larger initiative: How can we collect information about a user in a way that is non-invasive, productive, and that doesnt compromise the privacy of the user?

This problem can be partially solved with an intelligent UI that allows the user to constantly be presented with how information is tagged, and also allows the user to change and personalize that organizational scheme to suit himself. Furthermore, an automated process could be used that would systematically let the user go through his history, allowing him to tag and upload information about sites that he goes to often. A system like this could even dynamically create smarter bookmarks which take into consideration how often the user goes to specific sites as well as his searches and his tags.

4.2 Add in a social framework

Most of these services mentioned have a general anonymous body that tags information. This anonymity and independence should definitely be part of the solution. However. a real solution should come from a sort of double tagging: tagging your friends, communities, etc., and then tagging the information itself. By doing this we do not need to tag as rigorously, because each person will develop a profile that will in turn provide useful tagging information. More simply, adding tiers of social information mined from simple user choices and preferences can add extra searchable dimensions to these large sets of information and help provide better, more relevant content to users.

There are definitely privacy concerns here. The system would need to be developed to have simple means to control visibility of information. There should be tiers of information that are shared. Some information used only personally, some for friends, some for community, and some for the general population.

4.3 Metatags and visualization: hierarchies and clusters

If we intend to add more structure to the information provided by tagging systems, we must begin to implement ways to add hierarchy to the tags themselves. For instance “itunes,” “ipod,” “mac,” “powerbook,” could all be metatagged as “apple.” Allowing the user to use metatags for both groups of content, and groups of people, would allow for a much better understanding of specificity and context when searching for information. Some of this information can be gleaned from clustering algorithms; however, in that case there is greater difficulty in deducing which items are more general then others. Furthermore, these clusters and hierarchies should be easily viewable/editable by the user.

5 Conclusion

In summary, from analyzing the structure of metadata produced by tag-based systems in terms of information retrieval for the World Wide Web, we are presented with a variety of characteristics which can be taken advantage of to improve information retrieval, the most notable being the distribution of the link structure between tags. Furthermore, from critiquing the current state of web-based services which use tagging to organize a collective set of bookmarks, we see the opportunities for folksonomies to provide an organic organizational system for information. Both of the two popular social bookmarking tools, Del.icio.us and Yahoo! MyWeb 2.0, remain immature in their approximation of a full folksonomy. The addition of a social framework, as in the case of Yahoo! MyWeb 2.0, allows for more searchable dimensions, and is a necessary first step to extending the basic framework of tagging used in other more basic systems such as Del.icio.us. Finally, it is apparent that a more powerful and transparent means to collect user information is necessary, which also allows for more intricate aggregation schemes. Such a system should allow users to tag, in an intuitive way, all aspects of the system: content, users, and even tags themselves. A more flexible and powerful tagging system could better filter content to be personalized to an individual or tailored to a given subject matter.

References

[1] David Abrams. Human factors of personal web information spaces.

[2] David Abrams, Ron Baecker, and Mark Chignell. Information archiving with bookmarks: Personal web space construction and organization.

[3] Jose Borges and Mark Levene. Data mining of user navigation patterns.

[4] Allan Borodin, Gareth O. Roberts, Jeffrey S. Rosenthal, and Panayiotis Tsaparas. Finding authorities and hubs from link structures on the world wide web.

[5] Soumen Chakrabartia, Byron Doma, Prabhakar Raghavana, Sridhar Rajagopalana, David Gibsonb, and Jon Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text.

[6] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology.

[7] Scott A. Golder and Bernardo A. Huberman. The structure of collaborative tagging systems.

[8] Victor S. Grishchenko. Computational complexity of one reputation metric.

[9] http://collabrank.web.cse.unsw.edu.au/del.icio.us/. Collaborative rank.

[10] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective.

[11] Jon M. Kleinberg. Hubs, authorities, and communities.

[12] Ben Lund, Tony Hammond, Martin Flack, and Timo Hannay. Social bookmarking tools (ii): A case study - connotea. *D-Lib Magazine*.

[13] Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata.

[14] Nathalie Math and James R. Chen. User-centered indexing for adaptive information access.

[15] Alex G. Bchner Maurice D. Mulvenna, Sarabjot S. Anand. Personalization on the net using web mining: Introduction.

[16] Bill Raschen. A resilient,evolving resource: How to create a taxonomy.

[17] Terrell Russell. Contextual authority tagging: Cognitive authority through folksonomy.

[18] Christoph Schmitz. Towards self-organizing communities in peer-to-peer knowledge management.

[19] Clay Shirky. Ontology is overrated: Categories, links, and tags.

[20] Steven Strogatz and Duncan Watts. Collective dynamics of 'small-world' networks.

[21] Benjamin Szekely and Elias Torres. Ranking bookmarks and bistros: Intelligent community and folksonomy development.

[22] Thomas Vander Wal. Explaining and showing broad and narrow folksonomies – http://www.vanderwal.net.

[23] Jill Walker. Feral hypertext: When hypertext literature escapes control.

[24] Wikipedia. Definition: Folksonomy.

[25] Wikipedia. Definition: Information overload.

[26] Wikipedia. Definition: Pagerank.