## Title
Taggers versus Linkers: Comparing Tags and Anchor Text of Web Pages



## Authors
Liu, Yiming
Kumar, Ruchi
Lim, Kevin

## Publication Date
2008-04-01

# Taggers versus Linkers:
# Comparing Tags and Anchor Text of Web Pages

Yiming Liu, Ruchi Kumar, Kevin Lim
School of Information, UC Berkeley
{yliu, ruchi, k7lim}@ischool.berkeley.edu

UCB iSchool Report 2008-020
March 2008

Keywords: tagging, links, anchor text, Web, autotagger, characterization

**Abstract**

The Web is home to information services that generate vast quantities of content. This presents a tremendous challenge for services that organize this content and facilitate efficient and relevant retrieval. Tagging is one way to organize this information. Tags are short, user-selected keywords and phrases applied to a Web resource, which can serve as concise semantic descriptions of that resource. Hyperlink anchor text appears to serve similar purposes–namely, as user-generated, concise resource descriptions, useful for navigation and search. In this paper, we present a study characterizing tags and anchor text being used on the Web, with the motivation of implementing tag suggestion using existing Web metadata. We suggest that some degree of conceptual similarity exists between tags and anchor text, and that anchor text can contribute to tag recommendation. Further, it may be possible to extract semantic groups of tags based on existing usage, from which tags can be suggested. Finally, we assess how tags and anchor terms pertain to subtopics within a given document, and propose a window-based text processing method that can be used to discover subtopic tags.

## Contents

# 1 Introduction

The rise of 'folksonomies', or collaborative categorization system driven by social 'tagging', is one of the latest attempts to address the challenge of organizing, navigating, and retrieving the vast amount of information available on the Web. The prime examples are social bookmarking systems such as `del.icio.us`. In these systems, users post Web resource to the system and annotate them (web pages, images, videos, etc.) with tags–a set of brief, descriptive keywords and phrases. The social tagging system then aggregates these annotations across all users for each resource. Previous work [7, 10] have identified various types of tags and motivations for their use, such as descriptive tags for future retrieval, conceptual tags for clustering resources, or idiosyncratic tags for personal organization or opinion expression.

However, tagging is not necessarily a perfect solution for Web resource organization. Due to its free-form style, folksonomic schemes incur all the typical problems of an uncontrolled vocabulary [11]. Beyond even that, however, is a problem of sparsity. Similar to how many people store their own digital photos with remarkably undescriptive default filenames such as 'DSC00007.jpg', so do they leave a substantial number of bookmarks untagged. Thinking up tags is a barrier to entry.

To lower the barrier for tagging, systems such as `del.icio.us` provides a listing of recommended tags and popular tags when a user tries to post a new resource. However, in `del.icio.us`, these tag recommendations are only generated when a number of other users have already tagged the Web resource in question. Furthermore, recommendation effectiveness is directly dependent upon the quality of these previously applied tags. If previous users tagged a Web page merely 'Web', then the subsequent recommendations would also consist of the same tag and nothing more.

As the number of annotators increase, these problems tend to fade, as the number of users and the number of good tags tend to increase. However, given the scale of the Web and the resources it contains, it is unlikely that every resource would receive such attention. As folksonomic techniques are adopted within enterprise contexts, it is also unlikely that most corporate resources will be fully tagged for easy navigation and retrieval, given the typical size of enterprise document repositories relative to the number of employees.

At the same time, tagging is not the only way to assign meaning to Web content. By its fundamental design, the Web links its resources together via hyperlinks, which are often annotated with "highlighted clickable text" [5] called anchor text or link labels. Anchor text share some interestingly similar properties with tags. Anchor text associated with a hyperlink are usually brief keywords or phrases that describe the resource being referenced, and are often effective summaries of the resource [2, 5]. The author of a Web page may create a link to any other page, even if he does not control the destination page. As such, modern Web search engines aggregate anchor information when crawling the Web, and tend to give significant weight to anchor text information [2] to determine rank and relevance to search queries.[1]

This paper presents a preliminary study of the properties of tags and anchor text. User-generated tags and anchor text are compared at general levels, to characterize these two types of metadata with respect to each other. We then explore the possibility of using existing tags and anchor text for semantic grouping of documents. We examine how simple subtopic analysis can contribute to the identification of low frequency tags localized to specific regions in the document. These characterizations are designed as exploratory steps toward a solution for complementing social tagging systems with existing Web metadata.

# 2 Related work

Though a relatively new area, there has been some prior work in the analysis of social tagging and its characteristics. Anchor text, on the other hand, has been analyzed mostly under the lens of improving

[1] The practice of "Googlebombing", for example, involves many users linking to a particular site with a specific piece of anchor text (desirable or otherwise) to distort a search engine's view of the resource. See [1].

search engines and search techniques. Very little work has been done to compare these two types of metadata directly.

Of the small but growing literature in tags, Golder and Huberman [7] provide one of the most comprehensive characterizations of tagging behavior on `del.icio.us`. Of its many insights, their study finds that a majority of URIs reach a peak level of popularity (the rate at which this URI is being posted) within a few days of first being bookmarked, then stabilizing thereafter. After being bookmarked around 100 times, the combined tags of many users' bookmarks show stable patterns in which the proportions of each tag is nearly fixed.

Along a similar vein, Halpin et al. [8] define the "information value" of a tag to be the information conveyed by the natural language term used in the tag and how this distinguishes a particular resource from other resources. They modeled the information value of most frequent tags in a power law distribution by looking at their co-occurrence with other tags, and created inter-tag correlation graphs to chart their relation to one another. The authors suggested that this methodology could be developed further to possibly extract a formal classification scheme from a folksonomy.

Noll and Meinel [12] analyzed tag usage and compared the differences between authoritative metadata – that is, information provided by document authors – and collaboratively generated tags provided by readers of the same content. Among the findings, they suggest that users tend to focus on tagging popular pages. Of the Web documents in their collection, 52.3 percent of had low PageRank. These documents collectively received only 11.5 percent of all tags. They also found that 46 percent of tags occurred in author-provided content, which increased to 58 percent after lexical post-processing. Of the remainder, most were general terms that describe the resource's category rather than information about the resource itself.

Finally, Farooq et al. [6] characterized social bookmarking systems by proposing six tag metrics that expose properties of collections and those of individual tags. One of these metrics, tag non-obviousness, measures how often the tag itself appears in the content of the original resource being tagged. The implication is that a tag that does not appear in a document is "non-obvious", since a content-based method would not be able to identify it. They suggest that a non-obvious tag has more information value [8].

Earlier work also studied the properties of anchor link text. Brin and Page [2], in their description of Google, suggested that anchor texts can be effectively used for ranking hypertext documents, which they described as providing "more accurate descriptions of web pages than the pages themselves."

Similarly, Eiron and McCurley [5] characterized anchor text in context of page content and search query terms. They find that anchor text contain similar terms to titles in general, but contained terms that were relatively rare in the full-page content. When used in conjunction with search query terms, pages with consistent matching anchor text tend to be more relevant than title- or text content-based matching alone.

Chakrabarti et al [3] describe the notion of an "anchor window" to expand the radius of influence out from the actual anchor text. In one stance, using the Yahoo homepage URI as the target, the word '`Yahoo`' was typically within 50 bytes of the anchoring attribute (that is, the `href` attribute).

Davison [4] shows that anchor texts are not only similar to their referent pages, but also to link-peers of the document, which shows that anchors can point to a "topic locality" on the web. Also interesting is that an anchor is least similar to a random page from the corpus, as compared to titles and descriptions.

In all, the literature suggests that social tagging tends to favor popular, interesting items, leaving the remaining items to languish in relative obscurity. The problem may be self-perpetuating, as obscure items are not easily discovered, and its tag quality and quantity may remain relatively poor. A substantial proportion of tags can be discovered in the text of the resource itself, though it is unclear how they can be effectively discovered. The remainder are relatively non-obvious, and tend to be categorical descriptions.

Anchor text is typically used in search as a concise summary of the document to which it links, while even more descriptive context tends to lie in close proximity to the link itself. Such text can be closely aligned to authorial description, or at least document titles. They can potentially point to localized topics in a longer document, as per HTML's ability to refer to URI fragments.

The findings here inform our own work in the examination of tag and anchor properties, and motivate our investigation of tag grouping and subtopic tagging and linking.

# 3 Methodology

Since there is little prior work comparing user-generated tags to anchor text, our study conducts an initial comparative analysis, followed by more detailed exploration of interesting trends. A dataset was collected consisting of tags, the original textual contents associated with the tagged URIs, and the anchor text linking into those URIs. Using a custom-built web interface, tags and anchor texts were compared side by side for random given URIs. A comment system was included to store the qualitative observations. At the same time, more detailed analyses on tags and anchor text from the collection were conducted to compare various simple textual features of tags and anchor text.

From these observations and statistics, we identify two further measures of interest to develop. One metric involves assessing the quality and obviousness of tags, as opposed to anchor text, when compared against the content of the original Web document. The other involves the effectiveness of using either tags or anchor text in identifying specific subtopics within the original Web document.

The dataset was based around 11,019 URIs collected from `del.icio.us`. Using these URIs, we extracted all `del.icio.us` posts about the URI. Each post was a "bookmark" made by a `del.icio.us` user, and contained a title (usually the title of the Web page being bookmarked), a set of tags, a free-form annotation text, and a timestamp recording the time of posting. This provided the tagged document collection for our investigation.

We used the Technorati blog search engine as a source for inbound links and their associated anchor text. The major search engines also provided inbound-link retrieval, but their public interfaces are very restricted, and returned few or very poor and redundant inbound link results. Technorati proved to be a more useful source for anchor text, as blogs are often very link-oriented. Of the original 11,019 URIs, we were able to associate 6,649 URIs with anchor text.

We were able to retrieve the original content text from each of the URIs, and successfully done so for 10,328 URIs. These content-texts were hand-annotated with a label of "English" or "Non-English", to restrict ourselves to English texts for our analyses. 8,741 URIs were deemed English. Of these, 5,498 URIs were also associated with anchor text.

During the initial phase of the study, we constructed a web interface to view anchor text and tags, and their associated frequencies of occurrence, side-by-side for a given URI. This was attached to a commentary database, where qualitative observations could be recorded and reviewed at a later time.

# 4 Analysis

This section begins with a discussion of general qualitative and quantitative observations, to point out overall patterns and findings. Then, we report on two patterns of interest – the distribution of general vs. specific tags and anchor text, and the distribution of subtopic tags and anchors.

## 4.1 General comparisons

The initial comparison of tag and anchor text data was run in two phases. In the first, URIs were randomly selected, and we examined their associated tags and anchor text via our web interface. In the second, quantitative comparisons were performed to validate any interesting patterns that emerged in the first phase.

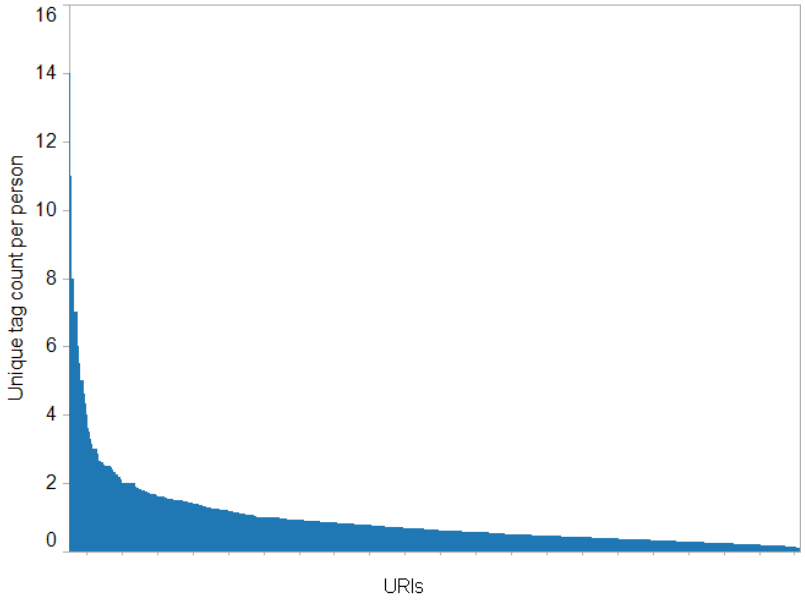


Figure 1: The distribution of frequencies of tags across the set of URIs, sorted by frequency and normalized by the number of users.

#### 4.1.1 Uniqueness and metadata diversity

Tags tended to be diverse, describing the resource, its general-level categories, its subtopics, and on occasion, personal commentary. Anchor text tended to be more direct descriptions for the resource itself, with minor lexical variation and the occasional commentary, rather than categories or higher-level semantic description.

Consider the example of Socialmedia.com, an advertising network based primarily on the Facebook social networking platform. The top four tags, by frequency, for this URI consisted of *facebook*, *web2.0*, *applications*, and *widgets*, which are indicative largely of the company's general category and business model. The top anchor text consisted of *social*, *media*, *social media*, or some variant thereof, interspersed with rare occurrences of *Facebook* and *advertising* and such less obvious labels that did not appear in the document content.

A quantitative validation was run for the entire dataset. Figure 1 shows the distribution of frequencies of unique tags across the set of URIs with five or more tags, sorted by frequency and normalized by the number of taggers. As expected, this resembles a classic power-law distribution, with a small number of resources with highly diverse set of tags and a long tail of resources with few tags. A follow-up analysis of the same dataset, normalizing by the total number of tags for that URI (so as to control for the fact that some items were simply poorly tagged), showed the same pattern.

For anchor text, shown in Figure 2, the distribution (again, normalized by number of linkers) retains the rough overall shape. A follow-up analysis, controlling for the total number of anchors for that URI, exhibits the same pattern.

While anchor text themselves are relatively direct and uniform, the text in immediate proximity to the anchors were more rich and descriptive. Using a 250-character proximity window (which is provided by Technorati), we find an assortment of unique n-grams around the anchor text, such as *Facebook*, *ad networks*,

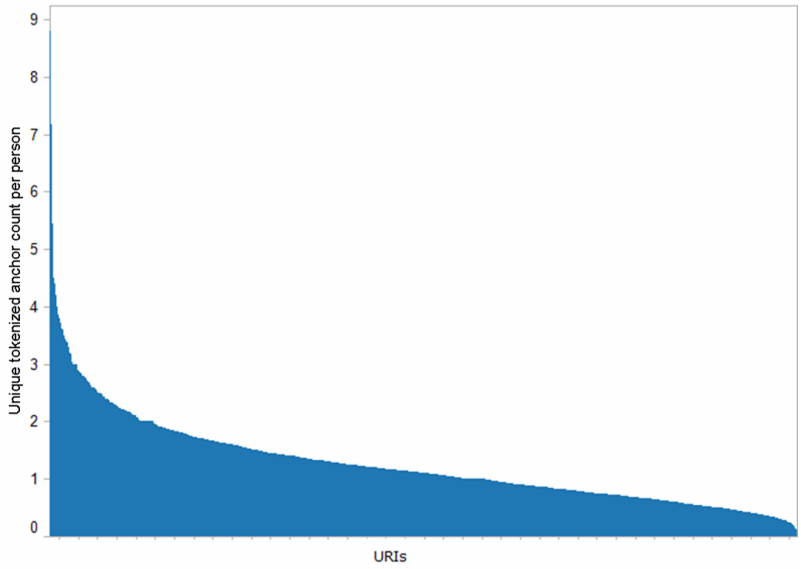


Figure 2: The distribution of frequencies of anchors across the set of URIs, sorted by frequency and normalized by the number of users.

*dollars*, and *Seth Greenstein*[2]. These provided a more useful and extensible view of the resource, in the same fashion – though perhaps still missing the higher-level semantic categories such as *Web 2.0*.

#### 4.1.2 Shared words and overlap

We examine tags and anchor words, with respect to each other and with the contents of the pages that they describe. Eiron and McCurley [5] found, via an overlap metric, that anchor text and page titles tended to be drawn from the same conceptual space of words. It remained to be seen whether tags and anchor text were likewise similar.

A simple lexical overlap metric can be used for assessing the similarity between tags $T$ and anchors $N$. For each URI, a set of unique tags was available. Further, anchor texts were tokenized, and a stoplist applied, to obtain a set of unique anchor words for the URI.

A simple overlap between the set of tags and the set of anchor tokens, is calculated by the simple ratio between the number of shared words over the total number of words:

$$overlap = \frac{\text{count}(T \cap N)}{\text{count}(T \cup N)} \tag{1}$$

Unfortunately, this approach yielded very small overlap values. Given the nature of the dataset, it is often the case that one set is much larger than the other. Even if all of $T$ existed in $T \cup N$, if $T$ is sufficiently small compared to $N$, the resulting ratio remained small and indistinguishable from no-overlap scenarios.

Instead, a "best of either" overlap comparison is more appropriate.

$$overlap = \max\left(\frac{\text{count}(T \cap N)}{\text{count}(T)}, \frac{\text{count}(T \cap N)}{\text{count}(N)}\right) \tag{2}$$

[2]Seth Greenstein is the founder of Socialmedia.com

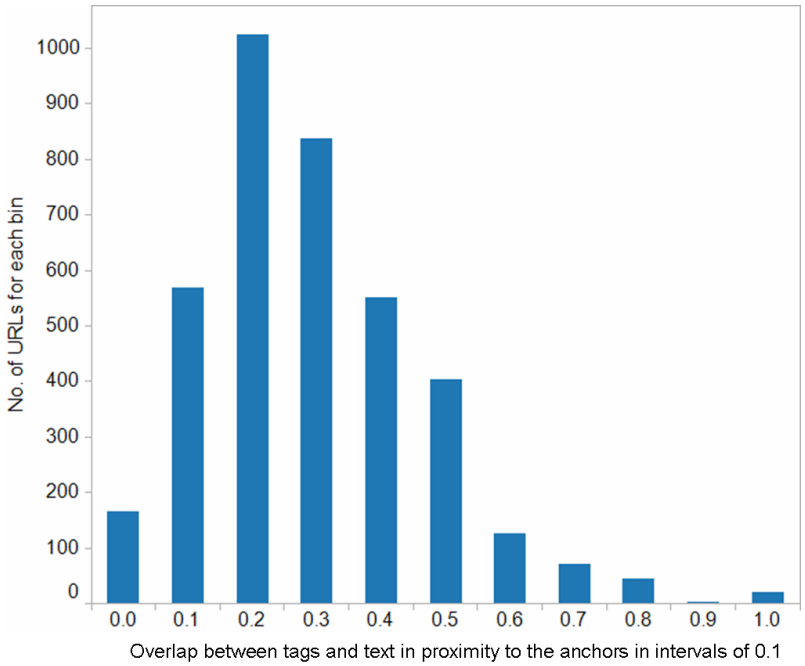


Figure 3: Overlap between tags and text in proximity to the anchor link, for each URI. Bottom axis is the overlap fraction.

This measure takes the better of two overlap scores from both tags relative to anchors, and vice versa. This value is used for comparison purposes.

Figure 3 shows the overlap between tags and words in close proximity to the link anchor. The mean fraction of overlap was 0.33, with a standard deviation of 0.16. Thus, there was some degree of overlap, but in many instances tags or anchor text introduced new words. It yet remains to be seen whether these new words are 'good' description words, especially in the anchor text case, for some value of 'good'.

#### 4.1.3 Tags, anchors, and titles

The same overlap process was run for tags/anchors relative to the page title. Prior observations indicated that anchors generally repeated or summarized page titles. While tags were less repetitive in that regard, some tags still tended to repeat keywords that were present in the title.

Figure 4 shows the overlap between tags and titles, compared with Figure 5 which shows the overlap between anchors and titles. As expected, the mean fraction of overlap between tags and titles was less than the overlap between anchors and titles. The mean fraction of overlap for tags was 0.384, while the same measure for anchors was 0.445.

The distributions share significant resemblance, except in a substantial number of cases where the entirety of the anchor text fully overlapped the title text, or vice versa. This is consistent with our previous observation that anchor texts tended to repeat title words more often than tags.

#### 4.1.4 tf.idf weights

Previous work (such as [8]) suggested that some tags might be obvious, by their overt presence in the original document or by their document-level distinguishing power. It is useful to assess relative importance of tag

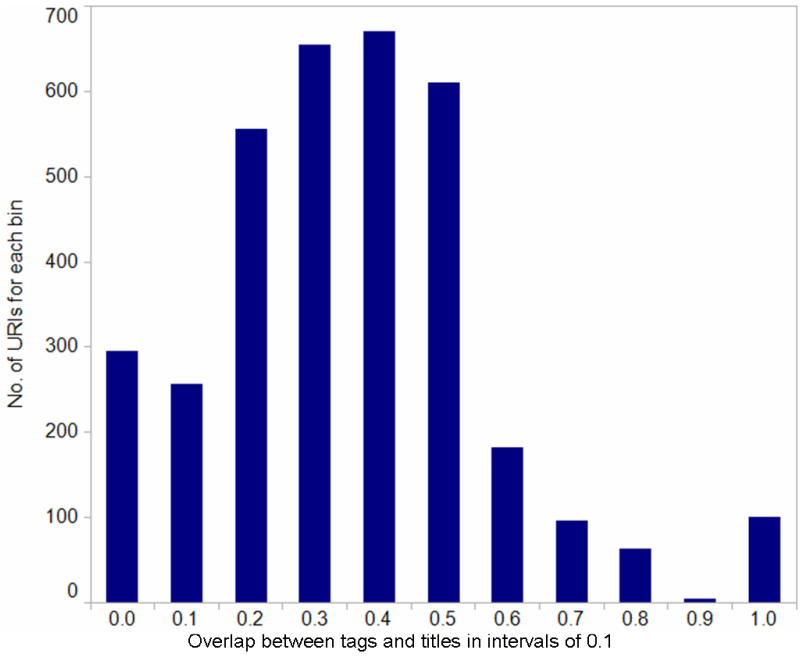


Figure 4: Overlap between tags and page titles, for each URI.

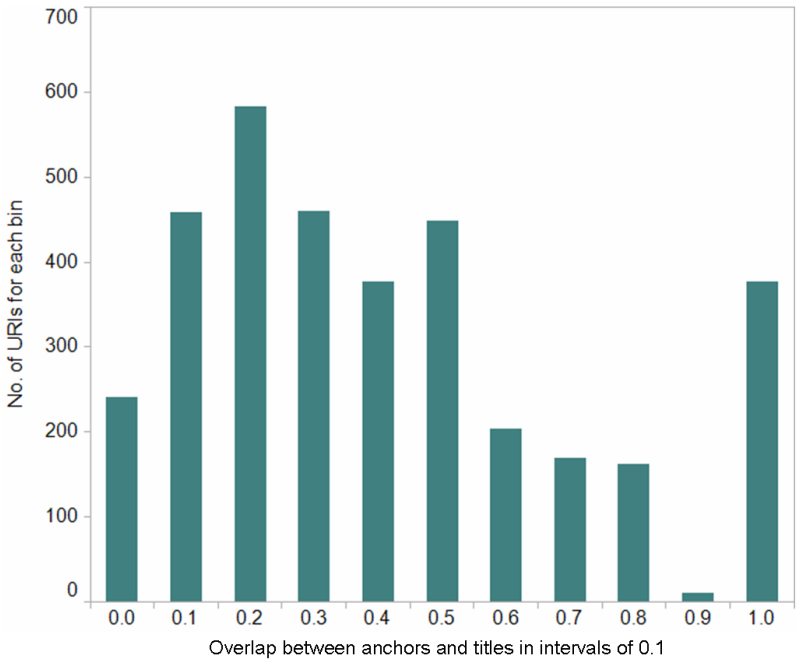


Figure 5: Overlap between anchors and page titles, for each URI.

and anchor words, in context of this distinguishing power.

A classical information retrieval measure for token weighting is tf.idf [13], which measures the relative importance of each word in a document relative to a collection. Using the full content dataset (i.e. all URIs for which we were able to retrieve full content) as the collection, tf.idf scores were computed and assigned to each token in a given document. Each tag and anchor token was checked against tf.idf weights (if these tokens, in fact, also occurred in the document).

A fair proportion of tags for a document did not actually occur in the document, as expected. In these cases, no tf.idf computation is possible.

The mean normalized tf.idf weight was 0.19 for tags, and 0.21 for anchor text. The average maximum tf.idf scores are approximately 0.32 and 0.43 respectively. Relatively speaking, thus, tags and anchor text tend to have fairly high tf.idf scores. On average, 56 percent of tags and 62 percent of tag or anchor text tokens that appear in the document will also be among the top five percentile of tf.idf-scored terms.

Of the high-tf.idf words that do not appear as tags or anchors, most are meaningless to humans when taken out of context by this bag-of-words model. For example, for the home page of Mouser Electronics, an online electronics shop, high-scoring tf.idf terms that were neither tags nor anchor text consisted of the words 'contact', 'alden', and 'optoelectronics'. 'contact' originates from the three 'contact us' links on the page, while 'alden' identified a Mouser supplier listed on the store front page. 'optoelectronics' referred to a product category being sold. Neither of the first two would have made good tags, though 'optoelectronics' might have been a useful tag for someone interested in such equipment.

Some tags and anchor text have good tf.idf scores, but tf.idf is not a foolproof predictor of tagworthiness. Some high-scoring tf.idf terms simply do not make sense as tags. It seems potentially difficult to separate, via automatic means, good tags terms from all high-tf.idf terms.

## 4.2 Tag/anchor categorization

Having reviewed the general issues in tag/anchor comparisons, we turn to more specific patterns of interest. One such issue involves assessment of categories for resources, via tag or anchor texts.

A basic problem of any kind of categorization systems is the granularity of terms or the levels of abstraction that they represent. Related terms describe an item along a spectrum ranging from very general to very specific. In tagging systems, a document tagged *ajax* or *python* may for be considered too specific for some and at the same time, too general by others who are more technically oriented. Indeed, prior work [12] suggests that of the percentage of tags not appearing in the content, many were simply categorical tags, occurring at varying levels of abstraction and granularity [11]. While the problem of establishing hierarchy is difficult and requires a well-developed ontological framework, we examine the possibility of identifying related groups of tags and anchor text.

### 4.2.1 General vs. specific tags

The tagset of a given document stabilizes over time following the Zipf distribution. Golder and Huberman [7] suggested that for such stable URIs, the more commonly used tags are at the general level in the above spectrum and have higher proportions whereas personally oriented tags are more varied and have lesser proportions.

As an example, Figure 6 shows the tag distribution for a sample URI titled, 'Using the FeedTools Cache in Plain Ruby Scripts'[3]. Above a certain frequency threshold, most tags tend to be general descriptions of the resource. Below the threshold, tags tend to be more related to subtopics in the resource or related domains to that resource. We noted that that the top few popular tags like *ruby* and *feedtools* describe the

[3] `http://dekstop.de/weblog/2005/12/feedtools_cache_in_ruby_scripts/`

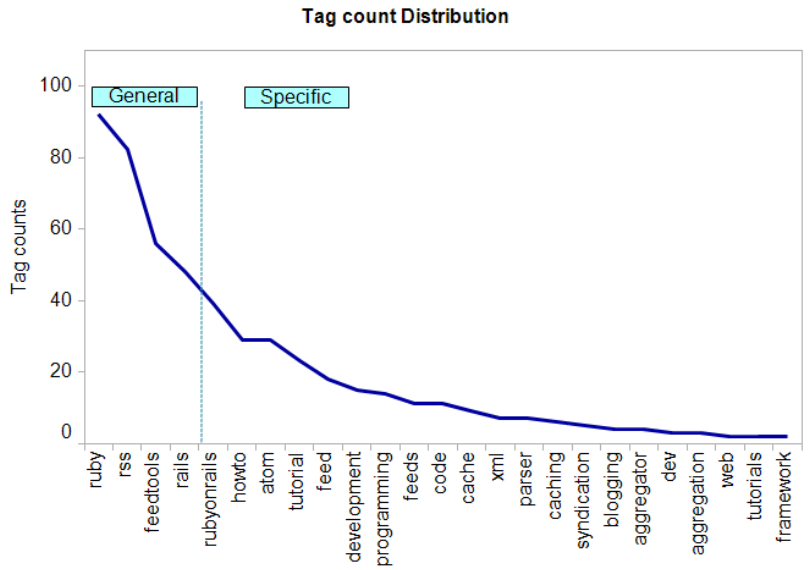


Figure 6: Tag distribution for a sample web page 'Using the FeedTools Cache in Plain Ruby Scripts'. Above a certain frequency threshold, most tags tend to be general descriptions of the resource. Below the threshold, tags tend to be more related to subtopics in the resource or related domains to that resource.

general level for this domain whereas less popular tags like *atom* or *parser* describe the more specific uses of the FeedTools library for Ruby.

### 4.2.2 Semantic groups

In order to gain insight into how semantic structures could emerge in distributed folksonomies, Halpin et al. [8] created inter-tag networks starting with a seed tag and finding other tags related to it by looking at direct or indirect dependencies of other tags on it.

Examination of only the tag terms, without their context, suffers from the problem of polysemy, as the same term can represent several meanings in different contexts. We explored this space differently by starting with a random seed document, and finding other documents similar to it using cosine similarity. The hypothesis is that a weighted bag-of-words, created by collecting the top tags for similar documents, should refer to a group description applicable to the set of these documents.

Cosine similarity between the seed URI and other documents in the collection was calculated using tf.idf scores normalized over the length of each document. We then looked at the tag-sets for the top ten most similar URIs to that of the seed URI. Since we were looking for only stable distributions, the URIs that had been tagged by less than 100 people [7] were ignored. For this set of similar URIs, a new bag-of-words was created consisting of the top non-distinct 50 percent tags, normalized over the length of the individual URI's tag-set. A limit of 50 percent was chosen as an arbitrary threshold. However, further work would be required to determine what cut-off percentage can be used.

Figure 7 shows the tag cloud for the bag-of-words created for the aforementioned seed document. This group gives a control over the problem of synonymy as related words appear in the same cloud.

The above procedure was repeated for other randomly selected URIs and the same pattern of semantic

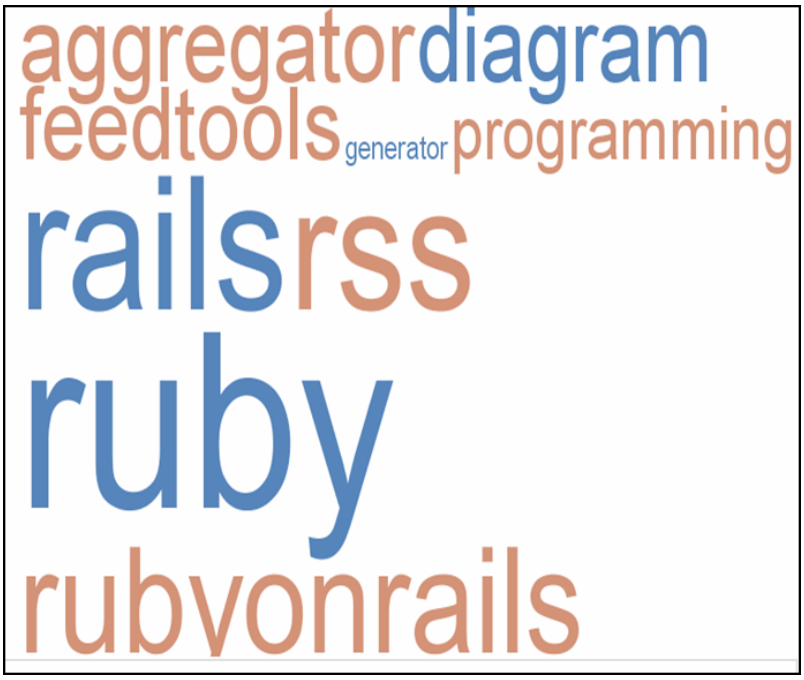


Figure 7: Tag cloud for tokenized bag-of-words representation for 'Using the FeedTools Cache in Plain Ruby Scripts'.

grouping was found to emerge. For example, one of the seed URIs [4] gave rise to the group related to *medicine*, and another one [5] to RSS (general group for the technical del.icio.us users).

To compare with anchor text, we examined the anchor text data for the same set of related URIs and created a tag cloud for the top 50 percent instances of non-tokenized, non-distinct anchor text. However, the URIs that we chose above for their stable tag sets did not consistently have a good amount of anchor text data. Using these sets, the same patterns were observable, but the relative sparsity of anchor text prevents stronger conclusions.

Based on the general pattern that anchor text tended to overlap with titles, we hypothesize that anchor text would tend to be resource-specific, rather than categorical; thus, in most instances, they would not likely exhibit the same pattern as tags. We leave more conclusive validation of this hypothesis to future work.

This approach to create semantic groups for given URIs can be a starting point for the emergence of automatic categorization schemes. This can further be useful for auto-recommending tags. One can recommend a popular tag from a related page that belongs to the same semantic group as the document being tagged.

## 4.3 Subtopics and tags

We aimed to find a simple operationalization of determining a non-obvious tag. Farooq et al. describe a given tag as non-obvious if it does not occur in the tagged document. However, that is not necessarily a comprehensive definition.

[4] http://medlineplus.gov/
[5] http://www.petefreitag.com/item/208.cfm

We ran a simple qualitative survey, consisting of graduate students at UC Berkeley, to generate thoughts on basic categories and properties of tags and anchors. Several respondents mentioned that some of the tags and anchor phrases assigned to a given URI did not pertain to the "whole" document, but instead described a specific subsection of the document. For instance, a large online listing of caffeinated beverages received a tag "starbucks" because many of the listings (seen close to one another in the alphabetical list) were Starbucks Coffee drinks. In another example, provided in Figure 8, a discussion of 'taste' in design occurs in a very localized area of the document.

This type of term, which only appears in a subsection of the document, characterizes a subtopic or a smaller subdomain within the document. These are tags that appear in the document, however, they appear with less frequency, occur in proximity and describe some specific information about the resource being tagged. They are also not necessarily obvious tags, when examined in context of the entire document. In a pure term-frequency count, they would probably by overshadowed by other terms. In the case of Figure 8, those seeking Oliver Reichenstein's view on design taste might indeed tag 'taste' as a term of interest.

We refer to these as "spike" or "less obvious" terms. We propose a method to determine these terms.

#### 4.3.1 Determining subtopic terms

One can intuitively think about an obvious term that appears regularly throughout the document, and contrast that with a term that only appears in a "spike." In the first case, this term is scattered around the full text, while in the second, it is concentrated in a cluster.

For each document, a window size in number of words is defined, which establishes a subset of the text at fixed size. The document, then, is a collection of contiguous subsets of the document of that fixed size.

A subtopic detection run sets the window at the beginning of the document and "slides" the window of focus across the document until it reached the end. The run examines the frequency of occurrence of this term in the window, relative to its occurrence over the entire document. If a term is localized to a small subset in a document – a spike term, it is likely to pertain to a subtopic. However, a term that is sprinkled throughout the document relatively evenly is likely to pertain to the whole document. This is the same basic intuition and algorithm used in TextTiling, which is a more advanced method for subtopic detection and segmentation in freeform text. [9].

**Example document:** "The roof, the roof, the roof is on fire. The rain, the rain, the rain, it falls in the plain." (20 tokens)

**Example window length:** 5 tokens

**Resulting Windows:**

1. "The roof, the roof, the"
2. "roof, the roof, the roof"
3. "the roof, the roof is"

...

15. "rain, it falls in the"
16. "it falls in the plain."

In the above example, the terms "roof" and "rain" are spike terms, while "the" is not.

To further quantify this notion, we prescribed a window score $W_s$ for each term $t$, and a given window length N in document D:

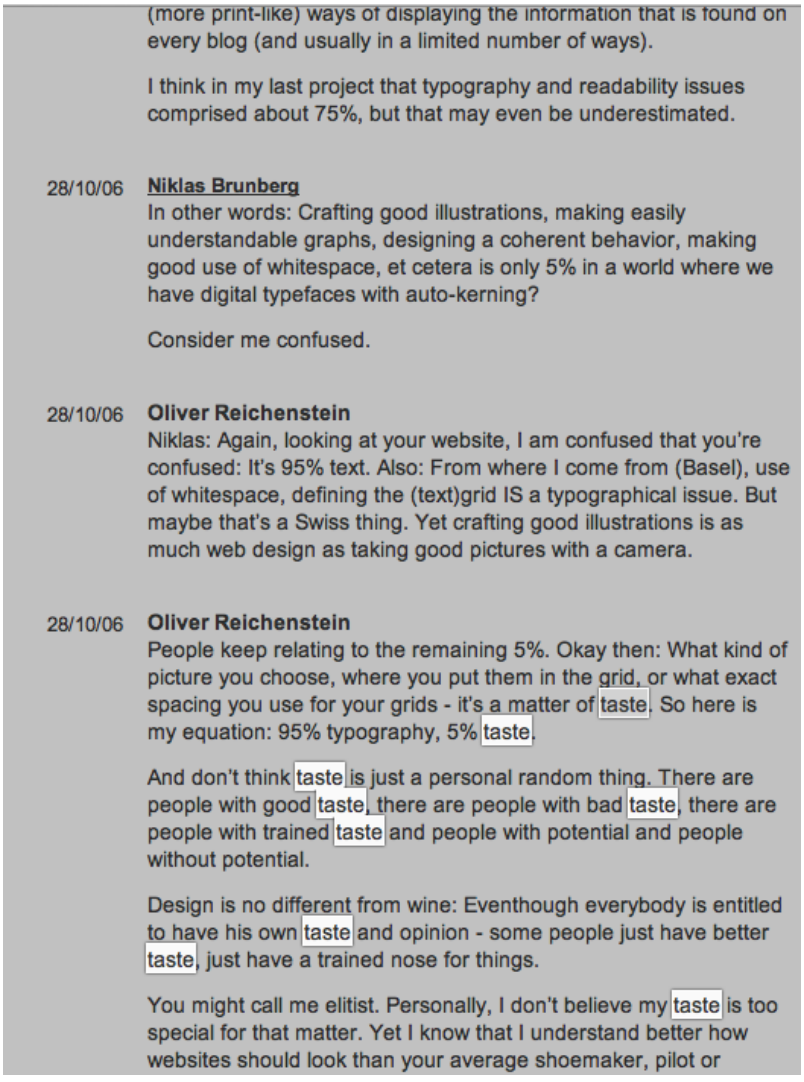


Figure 8: Observing a localized spike of a term in a document.

$$W_s = \frac{\max(T_w)}{T} \tag{3}$$

where $T_w$ is the occurrences of term $t$ in a given window subset, and $T$ is total occurrences of term $t$ in the document.

Spike terms in a given document will have window scores that approach 1. For example, the window score of term "roof" in the example document and sample window size five is calculated as 1, as all 5 of the occurrences of "roof" appear in a given window (the 2nd window listed). However, the window score for "the" will be 3/7, since the maximum occurrences in one window is three, but the document contains 7 "the"s.

#### 4.3.2 Subtopic Comparison of Anchors vs. Tags

We ran the window-based scoring on every tag and anchor term for every English document in our dataset. In this context, a fixed window did not make much sense as a window proportional to the size of the document– too small a window would not properly identify a spike, while too large a window would not provide enough locality. Experimental runs indicate that 10 percent of a given document was a fair window size.

Terms that only occurred in the document once are ignored, since they would all receive a window score of 1.

Results of these runs were promising. Tags and anchors have the same rate of subtopic description versus non-subtopic description. In particular, approximately 23 percent of tags and 20 percent of anchor terms presented strong subtopic scores in the 0.8 to 1 range.

This method was less successful in detecting very generic tags or anchors, accounting for less than 5 percent of very generic tags. We suspect that at such a level, these tags no longer appear in the document text.

While this method is used in an analytic context, to identify 'subtopic' tags and anchor terms, the same concept can be adapted (with some refinement) to determine subtopic tags. For usage as tags, terms that frequently occur locally may be as or even more interesting than terms that occur frequently at a document-global scale, as they point to specific topics covered in the document one might otherwise miss via titles or tables of contents.

### 4.4 Implications for auto-tagging

We note that a fair amount of overlap occurs between tags, anchors, and titles, but in our raw dataset, a substantial portion remains unique. It is unclear whether these terms would also make good tags. We have anecdotal suggestions to this effect, but no quantitative means to evaluate these suggestions. A user study would be effective here. If proven correct, this will have added a substantial amount of existing Web metadata, already known to be relevant to the resource in question, to the collection of words for tag suggestion.

Tags and anchors tend to have good tf.idf scores. However, it is dangerous to assume the contrary. It is entirely possible that words of high distinguishing power to an IR system are not necessarily "tag"-type words in the ways that humans think of tags. Furthermore, high tf.idf scoring terms are not comprehensive, limited by the amount of content available in the document.

Further, semantic grouping via document similarity analysis may suggest "related" tags, and provide some aid in resolving polysemy difficulties. The effectiveness of this technique can be easily tested in a deployed system.

Subtopic tag detection via window-based methods could yield both the upper-level semantic tags as well as the more special-interest tags used by select user groups. We think this to be a promising method for extracting 'spike' or 'less-obvious' tags from content, which a document-level extraction method may miss.

We believe that human evaluation will be useful for the creation of the final autotagging system. We envision the deployment of a system that would select some candidate set of tags based on the aforementioned set of data sources, and allow human judges to examine these recommendations to select the best set of tags. The effectiveness of the system can then be evaluated based on judges' selections.

# 5 Future work

As this is a first step into what appears to be an under-explored area, we leave much to future work. We note a few immediate steps to pursue.

First, it would be useful to develop a set of metrics for determining *quality* of tags and anchors. Based on qualitative observations, we suspect that the *quantity*, *uniqueness*, and *non-obviousness / less-obviousness* of tags and anchor text would be the primary factors in such a metric. A study using human evaluation would be able to validate some of our intuitions here. These metrics would also be useful in determining poorly described resources, which are prime targets for auto-tagging or tag recommendation systems.

Second, given the possible lexical variations in both tags and anchor text words, it may prove fruitful to apply a stemmer to these tokens and consolidate them before performing additional analyses. Stemming, however, increases the complexity of a system – aggressive morphological stemming may affect tag accuracy, while conservative stemming will not yield much benefit for the increase in complexity. Experimental runs will be required to verify the benefit of stemming techniques.

Finally, there is little substitute for implementation of an autotagging service and executing user tests on metrics of quality and usefulness.

# 6 Conclusions

We have presented a comparative analysis of user-generated tags and hyperlink anchor text, in context of annotating Web resources. A dataset of URIs, tags, anchors, and associated metadata has been compiled, and general analyses were performed on the metadata. We demonstrated that tags and anchor text shared similar traits, but also characterized a fair amount of differences, especially involving word diversity and semantic specificity. We also presented an approach to detect localized spike terms useful for subtopic detection.

We note that a significant amount of work remains to be done in this area. However, current results have been promising, and continued exploration of this area is warranted. We believe that our increased understanding of the nature and properties of tagging and linking will help in constructing automated tagging agents or tag recommendation systems, to assist in organizing and navigating the ever growing collection of pages and resources on the Web.

# References

[1] J. Bar-Ilan. Google Bombing from a Time Perspective. *Journal of Computer-Mediated Communication*, 12(3):910–938, 2007.

[2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[3] Soumen Chakrabarti, Byron Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon Kleinberg. Automatic resource compilation by analyzing hyperlink structure and

associated text. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 65–74, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[4] Brian D. Davison. Topical locality in the Web. In *SIGIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279, New York, NY, USA, 2000. ACM.

[5] Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 459–460, New York, NY, USA, 2003. ACM.

[6] Umer Farooq, Thomas G. Kannampallil, Yang Song, Craig H. Ganoe, John M. Carroll, and Lee Giles. Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In *GROUP '07: Proceedings of the 2007 international ACM conference on Conference on supporting group work*, pages 351–360, New York, NY, USA, 2007. ACM.

[7] Scott A. Golder and Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

[8] Harry Halpin, Valentin Robu, and Hana 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.

[9] Marti A. Hearst. TextTiling: A Quantitative Approach to Discourse. Technical report, Berkeley, CA, USA, 1993.

[10] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM.

[11] Adam Mathes. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. online, 2004.

[12] Michael G. Noll and Christoph Meinel. Authors vs. readers: a comparative study of document metadata and content in the www. In *DocEng '07: Proceedings of the 2007 ACM symposium on Document engineering*, pages 177–186, New York, NY, USA, 2007. ACM.

[13] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.