An Analysis of the Effectiveness of Tagging in Blogs

Christopher H. Brooks and **Nancy Montanez**
Computer Science Department
University of San Francisco
2130 Fulton St.
San Francisco, CA 94117-1080
{cbrooks, nmontane}@cs.usfca.edu

Areas of Interest: 05, 08, 11

Abstract

Tags have recently become popular as a means of annotating and organizing Web pages and blog entries. Advocates of tagging argue that the use of tags produces a 'folksonomy', a system in which the meaning of a tag is determined by its use among the community as a whole. We analyze the effectiveness of tags for classifying blog entries by gathering the top 350 tags from Technorati and measuring the similarity of all articles that share a tag. We find that tags are useful for grouping articles into broad categories, but less effective in indicating the particular content of an article. We then show that automatically extracting words deemed to be highly relevant can produce more focused categorization of articles. We also provide anecdotal evidence of some of tagging's weaknesses, and discuss future directions that could make tagging more effective as a tool for information organization and retrieval.

Introduction

In the past few years, weblogs (or, more colloquially, blogs) have emerged as a means of decentralized publishing; they have successfully combined the accessibility of the Web with an ease-of-use that has made it possible for large numbers of people to quickly and easily disseminate their opinions to a wide audience. Blogs have quickly developed a large and wide-reaching impact, from leaking the details of upcoming products, games, and TV shows to helping shape policy to influencing U.S. Presidential elections.

As with any new source of information, as more people begin blogging, tools are needed to help users organize and make sense of all of the blogs, bloggers and blog entries in the blogosphere (the most commonly-used term for the space of blogs as a whole). One recently-popular phenomenon in the blogosphere (and in the Web more generally) that addresses this issue has been the introduction of *tagging*. Tags are collections of keywords that are attached to blog entries, ostensibly to help describe the entry. While tagging has become very popular, and tags can be found on many popular blogs, there has not been (to our knowledge) much analysis devoted to the question of whether tags are an effective organizational tool, what functions tags are suited for, or the broader question of how tags can benefit users.

In this paper, we discuss some initial experiments that aim to determine what tasks are suitable for tags, how users are using tags, and whether tags are effective as an information retrieval mechanism. We examine blog entries indexed by Technorati and compare the similarity of articles that share tags to determine whether articles that have the same tags actually contain similar content. We compare this to clusters of randomly-selected articles and also to clusters of articles that share most-relevant keywords, as determined using TFIDF. We find that tagging seems to be most effective at placing articles into broad categories, but is less effective as a tool for indicating an article's specific content. We speculate that this is in part due to tags' relatively weak representational power, and conclude with a discussion of future work, focusing on increasing the expressivity of tags without losing their ease of use.

Background

Tags are keywords that can be assigned to a document or object as a simple form of metadata. Typically, users are not allowed to specify relations between tags. Instead, tags serve as a set of atomic symbols that are tied to a document.

The idea of tagging is not new; photo-organizing tools have had this for years, and HTML has had the ability to allow META keywords to describe a document since HTML 2.0 (Berners-Lee & Connolly 1996) in 1996. However, the idea of using tags to annotate entries recently become quite popular within the blogging community, with sites like Technorati [1] indexing blogs according to tags, and sites like Furl [2] and Delicious [3] providing users with the ability to assign tags to web pages and, most importantly, to share these tags with each other. Tags have also proved to be very popular in the photo-sharing community, with Flickr [4] being the most notable example.

This idea of sharing tags leads to a concept known as "folksonomy" (Shirky 2004; Quintarelli 2005), which is

[1] http://www.technorati.com
[2] http://www.furl.com
[3] http://del.icio.us
[4] http://www.flickr.com

intended to capture the notion that the proper usage of a tag is determined by the practicing community, as opposed to being decreed by a committee. Advocates of folksonomy argue that allowing the meaning of a tag to emerge through collective usage produces a more accurate meaning than if it was defined by a single person or body. Advocates of folksonomies as an organizational tool, such as Quintarelli (Quintarelli 2005), argue that, since the creation of content is decentralized, the description of that content should also be decentralized. They argue that centrally-defined, hierarchical classification schemes are too inflexible and rigid to be applied to the problem of classifying broad categories of documents such as Web data (in particular blogs), and that a better approach is to allow the "meaning" of a tag to be defined through its usage by the tagging community, which includes both bloggers and readers. This, it is argued, provides a degree of flexibility and fluidity that is not possible with an agreed-upon hierarchical structure, such as that provided by the Library of Congress' system for cataloging books.

To some extent, the idea of folksonomy (which is an argument for subjectivity in meaning that has existed in the linguistics community for years) is distinct from the particular choice of tags as a representational structure, although in practice the concepts are often conflated. There's no *a priori* reason why a folksonomy must consist entirely of a flat space of atomic symbols, but this point is typically contested by tagging advocates.

This discussion of tags and folksonomy highlights an interesting challenge to the traditional research community when studying subjects such as blogs: much of the discussion regarding the advantages and disadvantages of tags and folksonomy has taken place within the blogosphere, as opposed to within peer-reviewed conferences or journals. The blogosphere has the great advantage of allowing this discussion to happen quickly and provide a voice to all interested participants, but it also presents a difficult challenge to researchers in terms of properly evaluating and acknowledging contributions that have not been externally vetted. One goal of this paper is to move some of the discussion regarding tagging and folksonomies into the traditional academic publishing venues.

Tags, in the sense that they are used in Technorati and Delicious, are propositional entities; that is, they are symbols with no meaning in the context of the system apart from their relation to the documents they annotate. It is not possible in these systems to describe relationships between tags (such as 'opposite', 'similar', or 'superset') or to specify a 'meaning' for a tag, apart from the fact that it has been assigned to a group of articles. At first glance, this would seem like a very weak language for describing documents; users have no way to indicate that two tags are meant to be the same, or that a 'contains' relation exists between tags (for example, 'SanFrancisco' and 'California'), or that a set of tags form a complete enumeration of possible values (such as tags corresponding to the days of the week). Tagging advocates counter that these distinctions are too subtle for most users, who prefer a simpler, easier-to-use system, even at the cost of representational power, and that the collaborative determination of meaning makes tagging systems preferable to more powerful but less-widely-used systems (the RDF vision of the Semantic Web is typically offered as an alternative, often in a straw-man sort of way). Tagging advocates also argue that any externally-imposed set of hierarchical definitions will be too limiting for some users, and that the lack of structure provides users with the ability to use tags to fit their needs. This presents a question: what needs are tags well-suited to address?

In this paper, we focus on the issue of tags as a means of annotating and categorizing blog entries. Blog entries are a very different domain from photos or even webpages. Blog entries are more like "traditional" documents than webpages; they typically have a narrative structure, few hyperlinks, and a more "flat" organization, as opposed to web pages, which often contain navigational elements, external links, and other markup that can help to automatically extract information about a document's content or relevance. As a result, tags are potentially of great value to writers and readers of blogs.

About Technorati

Technorati [5] is a search engine and aggregation site that focuses on indexing and collecting all of the information in the blogosphere. Users can search for blog entries containing specific keywords, entries that reference particular URLs, or (most relevant to our work) entries that have been assigned specific tags.

We chose to use Technorati since they provide an open, freely-available RESTful (Fielding 2000) API. This provided us with programmatic access to their data, including the ability to find the top N tags, find all articles that have been assigned a particular tag, or all blogs that link to a particular URL.

Uses of Tags

We are particularly interested in determining what uses tags have. This question can take two forms: first, what tasks are tags well-suited for, and second, what tasks do users apply tags to? In this paper we will focus on the first question.

It is worth a brief digression to present our anecdotal observations about how tags are used empirically. There seem to be three basic strategies for tagging: annotating information for personal use, placing information into broadly defined categories, and annotating particular articles so as to describe their content.

Figure 1 contains a list of the top 250 tags used by blog writers to annotate their own entries, collected from Technorati on October 6, 2005. Examining this list immediately points out several challenges to users of tags and designers of tagging systems. First, there are a number of cases where synonyms, pluralization, or even misspelling has introduced the "same" tag twice. For example, "meme" and "memes", "Pasatiempos" and "Passatempos", or (more difficult to detect automatically) "Games" and "Juegos". It could be argued that a next-generation tagging system should help users avoid this sort of usage.

[5] http://www.technorati.com

About Me, Acne News, Actualite, Actualites, Actualites et politique, Advertising, Allmant, All Posts, amazon, Amigos, amor, Amusement, Anime, Announcements, Articles/News, Asides, Asterisk, audio, Babes, Babes On Flickr, Baby, Baseball, Blogging, Blogs, book, books, Business, Car, Car Insurance, Cars, category, Cell Phones, China, Cinma, Cine cinema, Comics, Computadores e a Internet, Computer, Computers, Computers and Internet, Computers en internet, Computing, CSS, Curiosidades, Current events, Data Recovery, days, Development, diario, Directory, Divertissement, Dogs, dreams, Entertainment, Entretenimento, Entretenimiento, Environment, etc, Europe, Event, EveryDay, Everything, F1, fAcTs, Family, fashion, Feeling, Feelings, FF11, FFXI, Film, Firefox, Flash, Flickr, Flutes, Food and Drink, Football, foreign-exchange, Foreign Exchange, Fotos, Friends, Fun, Funny, gnral, Game, Games, Gaming, Generale, General news, General Posting, General webmaster threads, Geral, Golf, Google, gossip, Hardware, Health and wellness, Health Insurance, History, hobbies, Hobby, Home, Humor, Hurricane Katrina, Info, Informtica e Internet, International, Internet, In The News, Intrattenimento, Java, jeux, Jewelry, jogos, Journal, Journalism, Juegos, kat-tun, Katrina, Knitting, Law, Legislation, libros, Life, Links, Live, Livres, Livros, London, Love, Love Poems, Lyrics, Msica, Macintosh, Marketing, MassCops Recent Topics, Me, Media, meme, memes, memo, metblogs, metroblogging, Military, Misc, Misc., miscellaneous, MobLog, Mood, Movie, Movies, murmur, Music, Musica, Musik, Musings, Musique, Muziek, My blog, Nature, News and politics, Notcias e poltica, Noticias y poltica, Opinion, Ordinateurs et Internet, Organizaes, Organizaciones, Organizations, others, Pasatiempos, Passatempos, PC, Pensamentos, Pensamientos, People, Personal, Philosophy, photo, Pictures, Podcast, Poem, poemas, Poesia, Poker, police headlines, Politik, Projects, Quotes, Radio, Ramblings, random, Randomness, Random thoughts, Rant, Real Estate, Recipes, reflexiones, reizen, Relationships, Research, Resources, Review, RO, RSS, Sade e bem-estar, Salud y bienestar, Sant et bien-łtre, School, Science, Search, Sex, sexy, Shopping, Site news, Society, software, Spam, Stories, stuff, Tech News, technology, Television, Terrorism, test, Tips, Tools, Travel, Updates, USA, Viagens, Viajes, Video, Videos, VoIP, Votes, Voyages, War, Weather, Weblog, Website, weight loss, Whatever, Windows, Wireless, wordpress, words, Work, World news, Writing

Figure 1: The 250 most popular tags on Technorati, as of October 6, 2005

In addition, many of the tags are not in English. This was an issue that we had not anticipated, but which is very significant to the experiments discussed below. Since we analyze document similarity using statistical estimates of word frequency, including non-English documents could potentially skew our results.

Finally, it seems clear that many users seem to use tags simply as a means to organize their own reading and browsing habits. This can be seen by the usage of tags such as "stuff", "Whatever", and "others." Looking at tags in Delicious produces similar results; along with tags indicating a web page's topic are tags such as "toRead", "interesting", and "todo". While this may be a fine use of tags from a user's point of view, it would seem to conflict with the idea of using tags to build a folksonomy; there's no shared meaning that can emerge out of a tag like "todo."

Figure 1 provides some evidence that many users seem to use tags as a way of broadly categorizing articles. This can be seen from the popularity of Technorati tags such as "Baseball", "Blogs", "Fashion", "Funny", and so on. While there is clearly great utility in being able to group blog entries into general categories, an open question remains: do tags provide users with the necessary descriptive power to successfully group articles into sets?

Finally, one of the greatest potential uses of tags is as a means for annotating particular articles and indicating their content. This is the particular usage of tags that we are interested in: providing a mechanism for the author of a blog entry to indicate "what a particular article is about." Looking at the list of most popular tags, it would appear that there are not many tags that focus specifically on the topic of the article. However, it may be the case that less popular tags are better at describing the subject of specific articles; we examine this hypothesis below.

Experiments

The primary question we were interested in addressing in this paper was how well tags served as a tool for clustering similar articles. In order to test this, we collected articles from Technorati and compared them at a syntactic level.

Dealing with non-English blogs

As mentioned above, one unanticipated wrinkle was that many of the blog entries are not in English. Since we analyzed document similarity based on weighted word frequency, it was important that non-English documents be removed, since we used an English-language corpus to estimate the general frequency of word occurrence. Our first naive approach was to use WordNet (Miller 1995) to determine whether a tag was a valid English word, and to discard documents with non-English tags. Unfortunately, that approach was ineffective, since many technical or blogging-related terms, such as "iPod", "blogging", "metroblogging", and "linux", are not in WordNet. Our current approach has been to construct an auxiliary "whitelist" of approximately 200 tags that are not in WordNet, but are in common usage in conjunction with English-language articles on Technorati.

We realize that this is only a stopgap measure - there are undoubtedly tags that are not on WordNet that are not on our whitelist, and it is certainly possible for someone to use an English tag to annotate an article written in Spanish. Our future work will include developing a classifier to recognize

non-English blog entries based on their text.

Experimental design

Our fundamental approach was to group documents into clusters and then compare the similarity of all documents within a cluster. Our hypothesis was that a cluster of documents that shared a tag should be more similar than a randomly constructed set of documents. As a benchmark, we also compared clusters of documents known to be similar. Finally, we constructed tags automatically by extracting relevant keywords, and used this to construct clusters of documents that shared statistically relevant keywords. This was intended to tell us whether humans did a better job of categorizing articles than automated techniques.

We began by collecting the 350 most popular tags from Technorati. For each tag, we then collected the 250 most recent articles that had been assigned this tag. For computational reasons, we restricted the experiment to 250 articles for each tag. HTML tags and stop words were removed, and a TFIDF (Salton & McGill 1983) score was computed for each remaining word for each article using the following formula:

$$TFIDF(word) = termFreq(word) * log(\frac{|corpus|}{DocFreq(word)})$$

Where $termFreq(word)$ indicates the number of times that word occurs in the blog article being processed. The second term in the formula provides an estimate of how common a word is in general usage. To compute this, a corpus of 8000 web pages were selected at random, and stop words and HTML tags removed. $DocFreq$ indicates how frequently a word appears in that corpus. This will cause commonly-used words to have a very low TFIDF score, and rare words to have a high TFIDF score.

Once we have a TFIDF score for each word in each article, we can construct clusters, one per tag, where a cluster contains a vector for each article bearing that tag.

For each cluster corresponding to a tag, we then computed the average pairwise cosine similarity (Baeza-Yates & Ribeiro-Neto 1999) of all articles in each cluster C, using the following equation:

$$aveSim(C) = \frac{\sum_{a,b \in C, a \neq b} cosSim(a,b)}{\sum_{i=1}^{|C|-1} i}$$

where

$$cosSim(A,B) = \frac{\sum_{w \in A \cup B} A[w]B[w]}{\sqrt{\sum_{v \in A} A[v] \sum_{v \in B} B[v]}}$$

The results of this experiment can be seen in Figure 2.

Figure 2 shows the rank order of tags on the x axis, with the most popular tag at the left, and cosine similarity on the y axis. As we can see, there is a small spike among very popular tags (centered around the tags “Votes”, “Games”, and “Game”). Apart from this peak, the similarity remains flat at 0.3. Interestingly, there is not an increase in similarity for

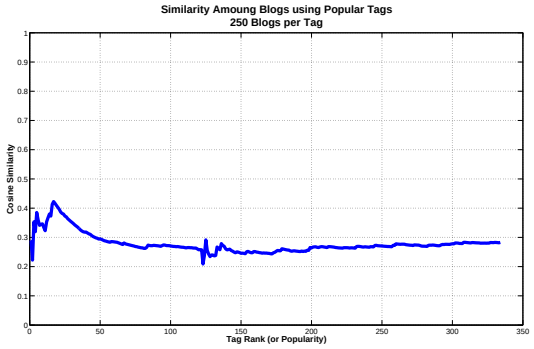

Figure 2: A Comparison of Tag Popularity versus pairwise cosine similarity

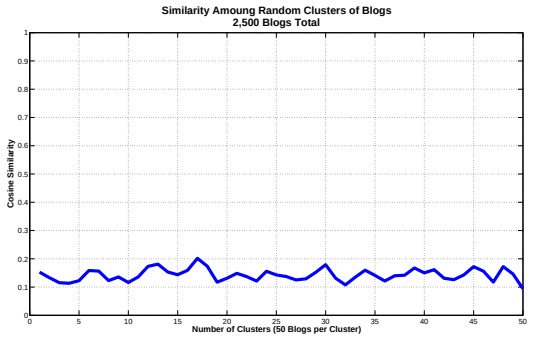

Figure 3: Pairwise similarity of randomly clustered articles.

rarely-used tags. Counter to our expectations, commonly-used tags and rarely-used tags seem to cluster articles at similar levels of effectiveness. We had expected that less popular tags would produce higher-similarity clusters, since these would likely be less-common words. In this experiment, that would appear not to be the case.

Taken in isolation, this is not very informative; does 0.3 indicate that a collection of articles are very similar, not at all similar, or very similar?

In order to provide a lower bound on the expected similarity measurements, we also conducted an experiment in which articles were placed into clusters at random, and the pairwise cosine similarity of these clusters was calculated. The results of this experiment can be seen in Figure 3.

As we can see from figure 3, the pairwise similarity of randomly-selected blog entries is between 0.1 and 0.2. This would seem to indicate that tags do provide some sort of clustering information. However, it is not clear whether an average pairwise similarity of 0.3 is a good score or not. We know that, if all articles are completely identical, the average pairwise similarity will be 1.0, but not how this score will decline for non-identical articles.

To address this question, we applied the same metric of average pairwise cosine similarity to articles grouped as “related” by Google News. The intent of this is to provide an upper bound by determining average pairwise cosine simi-

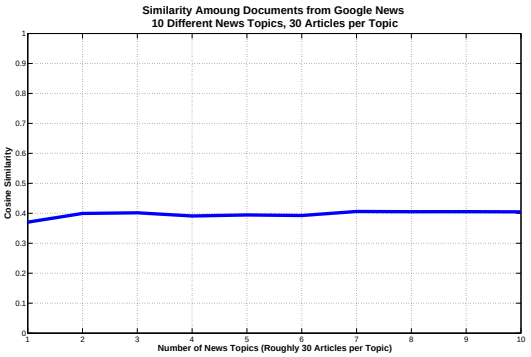

Figure 4: Pairwise similarity of clusters of articles deemed 'related' by Google News.

larity for articles judged to be similar by an external automated mechanism. The results of this experiment can be seen in Figure 4.

As we can see from Figure 4, articles classified as "related" by Google News have an average pairwise cosine similarity of approximately 0.4. Examining these articles by hand shows some articles that would be considered "very similar" by a human, and some articles that are more generally about the same topic, but different specifically. For example, under related articles about the nomination of Harriet Miers to the U.S. Supreme Court are specific articles about Bush's popularity, about Miers' appeal to Evangelical Christians, about cronyism in the Bush White House, and about Senator Rick Santorum's opinion of Miers. While these articles are broadly related, they clearly describe different specific topics, and we would not expect them to have a pairwise cosine similarity of anywhere near 1.0.

We can conclude from this that tagging does manage to group articles into categories, but that there is room for improvement. It seems to perform less well than Google News' automated techniques.

Automated tagging

As a first step towards providing tools that will assist users in effectively tagging articles, we tested the similarity of articles that contained similar keywords.

We selected 500 of the articles collected from Technorati and, for each of these articles, we extracted the three words with the top TFIDF score. These words were then treated as the article's "autotags." We then clustered together all articles that shared an autotag, and measured the average pairwise cosine similarity of these clusters. The results of this experiment are shown in Figure 5.

Interestingly, simply extracting the top three TFIDF-scored words and using them as tags produces significantly better similarity scores than tagging does (or than our evaluation of Google News, for that matter). The clusters themselves are typically smaller, indicating that automated tagging produces more focused, topical clusters, whereas human-assigned tags produce broad categories.

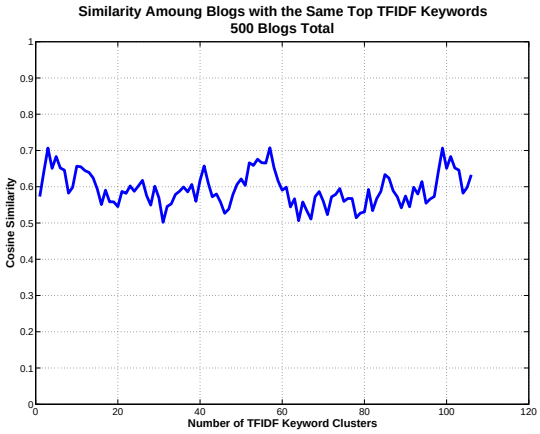

Figure 5: Pairwise similarity of clusters of articles sharing a highly-scored word.

Discussion

The experiments discussed above raise a number of interesting questions. The first is whether the metric we have used is in fact an accurate way to measure the similarity of blog entries. This is a completely syntactic measure, based primarily on frequencies of single words occurring. Sentences are not parsed or analyzed, synonyms are not detected, and larger phrases are not looked for. These are all potential future directions; however, what is really needed is a means to calibrate our similarity metric, by measuring its performance relative to larger sets of articles deemed to be similar by an external source.

Assuming that our similarity metric has some merit, this experiment sheds some light on what it is that tags actually help users to do. These results imply that tags help users group their blog entries into broad categories. In retrospect, this is not a huge surprise; given that tags are propositional entities, we know that a tag's expressive power is limited to indicating whether or not an article is a member of a set. Additionally, since tags cannot be combined or related to each other, users must create a new tag for each concept they wish to assign to a blog entry. It is not hard to imagine that most users would not want to create a vast number of unrelated tags; rather, they would choose a smaller, more general set. These experiments lend credence to this theory.

Figure 1 also provides an interesting snapshot of how tags are used by groups of users. At least within this picture, it would seem that bloggers are not settling on common, decentralized meanings for tags; rather, they are often independently choosing distinct tags to refer to the same concepts. Whether or not the meanings of these distinct tags will eventually converge is an open question.

Future Directions

While it seems clear that tagging is a popular and useful way for bloggers to organize and discover information, it also seems clear that there is room for improvement.

For one, we maintain that a more expressive representation for tags is needed. Observing the way that tags are used in, for example, Delicious, it is apparent that newcomers want to use phrases, rather than single words, to describe documents. Eventually, users realize that phrases are not effective and construct multi-word tags, such as "SanFranciscoCalifornia." Unfortunately, these multi-word tags are also inflexible; there is no way to relate "SanFranciscoCalifornia" to either "San Francisco" or "California" in current tag systems.

In particular, we argue that users should be able to cluster tags (i.e. 'Baghdad', 'Tikrit', and 'Basra' tags might be contained within an 'Iraq' cluster) to specify relations (not just similarity) between tags, to use tags to associate documents with objects such as people. This is not necessarily inconsistent with the idea of folksonomy; there is no reason why hierarchical definitions can't emerge from common usage.

One of tagging's biggest appeals is its simplicity and ease of use. Novices can understand the concept (although they may try to use phrases rather than isolated symbols). Tags are easy for authors to assign to an article. The importance of this cannot be overstated; any extensions to current tagging systems must retain this ease of use. Complex languages or cumbersome interfaces will mean that tags simply will not be used. We plan to incrementally develop tools that allow users to cluster their tags in a low-impact, easy-to-understand way, automating as much of the work as possible.

We also feel that tools that can help users automatically tag articles will be of great use. In fact, one might argue that even the act of manually assigning tags to articles is too much burden for the user, as it forces her to interrupt her writing or browsing to select appropriate tags. We plan to develop extensions to the approach described above that automatically extract relevant keywords and suggest them as tags. Additionally, this tool should interface with social tagging systems such as Technorati and Delicious to determine how these suggested tags are being used in the folksonomy. It should detect if there are synonymous tags that might be more effective, and assist users in assigning tags in a consistent manner.

Finally, we are also interested in the evolution of tags as a social phenomenon. Tagging is an interesting real-world experiment in the evolution of a simple language. Anecdotally, we have seen that some tags, such as "linux" or "iPod" have relatively fixed meanings, whereas other tags, such as "katrina", have a usage that varies widely over time. In June 2005, "katrina" would most likely be associated with articles about a woman with that name. In the weeks following Hurricane Katrina, it would likely be associated with articles about the damage to New Orleans, and presently it might be associated with articles that describe the political fallout of the disaster. We plan to study this evolution more systematically, repeatedly collecting the top tags from Technorati and Delicious and comparing the articles they are tagged with to look for drifts in meaning.

Acknowledgments

We would like to thank Technorati for providing us with access to their data.

References

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. New York: Addison-Wesley.

Berners-Lee, T., and Connolly, D. 1996. Hypertext markup language specification – 2.0. Technical Report RFC 1866, MIT/W3C.

Fielding, R. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. Dissertation, University of California, Irvine.

Miller, G. A. 1995. Wordnet: A lexical database for english. *Communications of the ACM* 38(11):39–41.

Quintarelli, E. 2005. Folksonomies: power to the people. Paper presented at the ISKO Italy-UniMIB meeting. Available at http://www.iskoi.org/doc/folksonomies.htm.

Salton, G., and McGill, M. J. 1983. *An Introduction to Modern Information Retrieval*. New York: McGraw-Hill, Inc.

Shirky, C. 2004. Folksonomy. Blog entry at http://www.corante.com/many/archives/2004/08/25/folksonomy.php.