Smarter Blogroll: An Exploration of Social Topic Extraction for Manageable Blogrolls

Eric Baumer – Department of Informatics, University of California, Irvine

Danyel Fisher – Community Technologies Group, Microsoft Research

Abstract

The already huge number of blogs in existence is increasing rapidly, and many users are struggling to find a way to keep up with the expansion. A number of existing tools aim to capture the general topics of all currently popular topics among the entire blogosphere, while others allow individuals to read a fixed list of blogs. However, few personalized tools exist to help the individual get an overview of the specific blogs he or she reads. This paper presents the concept of social topic extraction via the Smarter Blogroll, which displays current topics for a selection of blogs. While there was little difference in users' ability to identify topics using the Smarter Blogroll, the results of our user study point to design recommendations for improving the use of metadata within blogroll entries, thereby facilitating blog reading. The paper concludes with implications for the design of tools to aid in blog navigation and reading, as well as recommended directions for future research.

Introduction

Ambient chatter. Gossip. "Hot" topics. Having one's finger on the "pulse." Communities of various sorts—social groups, communities of practice, and other types—all have sets of current interests. Getting to know the "pulse" of a community of any sort can be a challenge. In the case of blogs, technologies such as RSS and practices such as live-blogging mean that blog authors and readers can keep up-to-date with

happenings in their communities 24 hours a day. However, there is no current blogging equivalent to "ambient chatter" whereby salient topics across a community of blogs may easily catch the reader's eye. Furthermore, for some bloggers, keeping up with their communities is an ever more difficult task: according to Technorati's State of the Blogosphere report from August 2006 [15], the size of the blogosphere has doubled consistently approximately every 200 days, with the number of blogs at the time totaling over 50 million.

There are some current technologies that attempt to distill the topics of interest to the entire blogosphere or the internet as a whole. Technorati, among others, offers a "tag cloud" of the currently popular categorical tags across all blogs. However, 50 million voices blogging at once quickly changes from ambient chatter to the roar of the crowd, which does not help give a good impression about what is currently interesting to a specific community. When the scope is more limited than the entire blogosphere, topic extraction methods have been used for tasks such as automatic summarization of news articles [12] or consumer product reviews [2]. Topic extraction has also been used to summarize personal information, such as email [14]. Rather than limiting the amount of information one gets, as through summarization, other recent projects have emphasized instead limiting the scope of the search: del.icio.us stores an individual's favorite links, and CiteULike [3] stores their favorite articles. Both "Rollyo" and "Windows Live Search Macros" create scoped searches, allowing a user to limit their interests to a subset of the web.

"Blogrolls" — public lists of other blogs that a blogger reads — is one way of explicitly defining members of a blogging community. However, blogrolls also serve a number of other functions. They are a way of socially scoping one's interests by focusing on a trusted list of sources. Displaying a blogroll is a way of both showing off the blogs one is interested in, and a way of vouching for those blogs [8]. Yet seeing a simple list of blogs can be challenging for a visitor: the name alone is often not enough to figure out what a blog is about. It takes a practiced eye to know what interest is shared by Eschaton, Apartment 11D, and Crooked Timber. (All three are political discussion blogs.)

In this paper, we present the Smarter Blogroll, which uses text mining techniques to augment a blogroll with information about the current topics of the blogs in that roll. By doing so, it serves several functions: first, it allows the owner of the blogroll to quickly and easily see at a glance what the current topics of discussion in their blogs are. It also allows visitors to get a general sense of the topics of the blogroll, and to understand what links might be most relevant.

This paper argues that combining topic extraction with social search can provide a rich experience of navigation in blogs. Rather than focusing, on the particular advantages of any one topic analysis algorithm, it presents a novel approach to designing tools for members of a blogging community.

The Smarter Blogroll was designed to meet a number of requirements.

Tasks

- A user should be *able to see what his or her blogs are talking about* at a glance.
- It should be *easy to find posts about a topic of interest*, both from the blog in which the post was found and on other blogs that happen to be discussing the same topic.

Design Features

- Since the Smarter Blogroll is meant to replace the current blogrolls that many bloggers incorporate on their blog, it should provide *at least the same functionality as current blogrolls*.
- The interface should also be such that it *visually resembles current blogrolls* and fits into roughly the same form factor on the screen.
- Because the Smarter Blogroll is meant to be a peripheral tool rather than a central one, its interface should be readily comprehensible and its functionality should be *generally simple*.

Design Philosophy

The design of the Smarter Blogroll draws on several different trends in thinking about blogs and social search to present a novel reader-centered approach to blogging tools.

Readers

Most previous blogging studies have focused on blog authors and tools for authors, as noted by Nardi et al., who predict that "future research is sure to pay attention to blog readers" [10]. This has left a gap in the literature: little has been published describing how blogs are read, or providing readers with tools to help them read blogs. Commercial utilities like RSS

aggregators exist to assist readers, but these tools simply collect and present content without aiding the user in actually reading or approaching the content. One of the few exceptions is NusEye [4], a blog reading tool that combined social network analysis with content analysis to present an alternative view for users of webfeed aggregators.

The work presented here represents a first step in exploring the creation of tools to assist users in the task of reading blogs. As Nardi et al. [10] note, readers are just as important a part of a blogging community as are authors, since readers and authors both share in the co-construction of the blog. Furthermore, because blogging is a complex social process in which the authors of one blog are often the readers of many others, this research explores an approach to developing tools that flexibly accommodate both authors and readers.

Blogrolls

A blogroll is a list of links to other blogs that can be found on the side of many blogs. Although many studies mention blogrolls, few have focused on the way they are used. This is surprising, as blogrolls are common in blogs: in Herring's genre analysis of blogs [5], over half of the 203 blogs studied linked to other blogs. Schiano et al. mention the importance of finding blogs through links on other blogs [13], but they focus their analysis on the content of the posts. A number of researchers have performed social network analyses based on blogs [6, 9], but these studies focus on the connections between blogs rather than how authors and readers actually use blogrolls.

Ostensibly (and as discussed by [10, 13]), a blogroll is a list of blogs that an author reads. That said, Lenhart's findings [8] suggest that there is no single consensus on the use for a blogroll, but rather a wide range of uses: in different blogs, blogrolls can be a declaration of interests, a list of social ties, an imprimatur of approval, or a way of referring users to related material. This broad array of functions suggests that blogrolls are an important part of a blog, both for the author and for the readers who visit it. Yet blogrolls can grow long and hard to read: our informal review of blogs showed many blogs with over fifty or one hundred blogs on their blogrolls; few blogrolls were broken into categories, and none were annotated with comments or notes. This leaves a challenge for readers in determining which blogs to read. It is hard to know what the topic of a blog is when many blogs have cryptic names. The only way to decide whether to read a linked blog all-too-often is to click through and read it, a prohibitively time-consuming approach. This means that the blogroll is

both less able to refer readers to related material and less able to show off the blogger's interests. Our goal is to alleviate this opacity and allow readers to know something about the blogs that the blogroll shows.

The Smarter Blogroll seeks to assist in the use of the blogroll as a means of communication among members of a blogging community, particularly between authors and readers. Specifically, we look to facilitate two of these uses for blogrolls: referring readers to related material, and displaying the author's interests.

Social Search and Social Browse

A blogroll can serve as a starting point for social browsing. Social browsing is analogous to social search: where social search means seeking out information by consulting one's local social network, social browsing refers to scanning through references collected by a set of trusted individuals. Search and browse are at opposite ends of a continuum of information finding behavior, of course; the relevant difference is that in a browsing scenario, the user does not necessarily have a strong, specific information need.

In order to better understand the possibilities, we can parameterize the design space in the following way. On one hand, is the task being supported a browse task or a search task? On the other hand, is the information presented being scoped? Here, we are specifically interested in community-based scoping, that is, scoping to a list of trusted resources within one's community. This parameterization is articulated in Table 1 with examples that pertain to blogging. For unscoped searching, Technorati provides a search of the entire blogosphere. For scoped searching, individuals can use tools like Rollyo or macro searches to restrict search engines to specific domains or websites, such as certain blogs of interest. While these custom search tools are not specifically designed for blogs, they still allow the user to search his or her blogroll. For unscoped browsing, Technorati provides a tag cloud that visually represents what the most popular tags are within the blogosphere. However, there are currently few if any widely available tools that support the scoped browse within blogs. This is the portion of the design space in which we are interested and which this paper addresses.

Table 1. Parametirizing the design space of tools to support blog reading.

	Unscoped	Scoped
Search	Technorati search	Rollyo, macro search
Browse	Technorati Tag Cloud	Blogroll

RSS readers are one of the few tools that support browsing by gathering disparate bits of content into a single location, but they do little to actually facilitate the browse task. RSS aggregators do not provide any information to the user beyond the feed contents: that is, they may reduce the number of clicks, but they neither provide any metadata nor enhance the reading experience. This paper seeks to explore the facilitation of the social browse task through the use of a tool that augments the blogroll with the "ambient chatter" of current conversations within the community, facilitating social browsing and enabling the user to easily "overhear" a discussion on another blog that might be of interest.

User Scenario

In order to contextualize our approach, we consider several common scenarios that help situate potential uses of tools like the Smarter Blogroll.

Monitoring My Own List

Martha has a few minutes before her next meeting, and decides to skim through her favorite blogs. She goes to her own blog, using the blogroll as a series of bookmarks. She wants to get a general sense of what interesting things are being discussed on her blogroll today. Usually, she would have to click on each blog, one at a time, skim over the blog, and then go back. She would get a chance to see some topics, but not many, and might miss out on blogs that are fairly far down her list.

With a dynamic, topic-oriented blogroll, Martha can see what the top topics are at a glance. She sees that several different blogs have mentioned "sock puppet" as a current topic. Apparently, there is news about sock puppets—or just an internet meme floating around. She follows a few of the links, and reminds herself to check it out later as she attends her meeting.

Understanding Another's List

Bob is interested in social networking and has stumbled across a relevant post on "Crooked Timber". He reads a few adjacent posts, and realizes that he is seeing more discussion of foreign policy and social networks then he has seen before. Bob checks the Crooked Timber blogroll to try to figure out where to look next. Unfortunately, the blogs on the list have rather confusing names: should he check "Ampersand", or "Fafblog?" Bob begins the process of clicking on each blog. After looking at a few, he finds nothing relevant and shortly gives up.

What Bob needs is additional metadata about each blog. A list of recent posts, or categories, might be valuable; so, too, might be some recent topics. With our topic-oriented blogroll, for example, Bob can quickly skim through the set, checking whether any seem interesting. On his first, accidental, visit to Crooked Timber, Bob now has a quick summary of the topics that interest the Crooked Timber blog community.

Implementation

In order to explore some of these concepts, we have built a tool called the Smarter Blogroll. In this section, we present the user interface of the tool, describe the client-server design, and lastly discuss our choice of topic extraction methods.

User Interface

The Smarter Blogroll is implemented as a Javascript front-end, and can so be embedded on either the sidebar of a user's blog, or on its own page. (The Javascript functionality is very similar to the Javascript blogrolls run by other blogroll services, e.g., blogrolling.com). It is implemented using Microsoft's live.com Gadget API, which provides convenient access to dynamic HTML functions and network access, and can be embedded on both live.com and third-party web sites, thus allowing it to function similarly to current blogrolls.

The Smarter Blogroll presents two views to a user: a blog-oriented view, which starts with a list of blogs (Figure 1), and a topic-oriented view, which starts with a list of topics (Figure 2). Both of these associate blog posts with automatically-extracted topics; users navigate through topics to get to the relevant posts on the blogs.

The blog-oriented view closely resembles a classic blogroll: it presents a vertical list of the blogs on this blogroll. However, this blogroll is annotated with metadata: each entry has up to five "top current topics;" each topic, in turn, unrolls on a click to show recent posts on that topic. For example, the blogroll in Figure 1 includes the blog Alabama Improper, whose current top topics are "huffington post," "daily kos," "mystery guest blogger," "little miss sunshine," and "xm radio." Automatically extracted

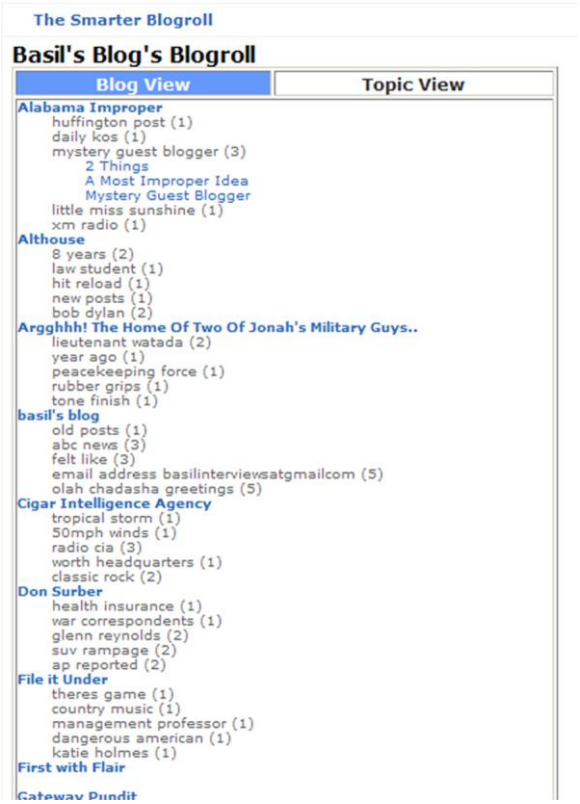

Fig. 1. The Smarter Blogroll, in Blog View. The user has selected the topic "mystery guest blogger."

topic entries are ordered by their relevance scores (see Eq. 2): higher-scoring entries are placed near the top of the list. These topic phrases serve to provide the "ambient chatter" for the blogroll: just as a listener's ear might take note when an uncommon phrase is overheard, a high-scoring topical phrase rises to the top of the topic list to catch the user's eye.

Note that some blogs, such as First with Flair in Figure 1, do not have any topics listed. There are a number of cases in which this might happen: the blog may not have any current posts; the topic extraction algorithm may not find any phrases that represent topics; the RSS feed for the blog may not contain any actual content. In this case, although the user does not have the added benefit of automatically extracted topics, he or she may still click on the name of the blog to visit the blog's front page.

The topic-oriented view (Figure 2) gives an overall impression of the topics being discussed across all the blogs on the blogroll. As with topics for individual blogs, these topics are ranked by relevance score (see Eq. 2), with all topics whose score is above a fixed threshold labeled top topics. Here as well, a number is listed in parentheses next to a topic to indicate the number of posts about that topic. By clicking on the topic, the view expands to show those posts. As in the blog view, the user can click on the post title to be taken to the permalink for the post, multiple topics can be

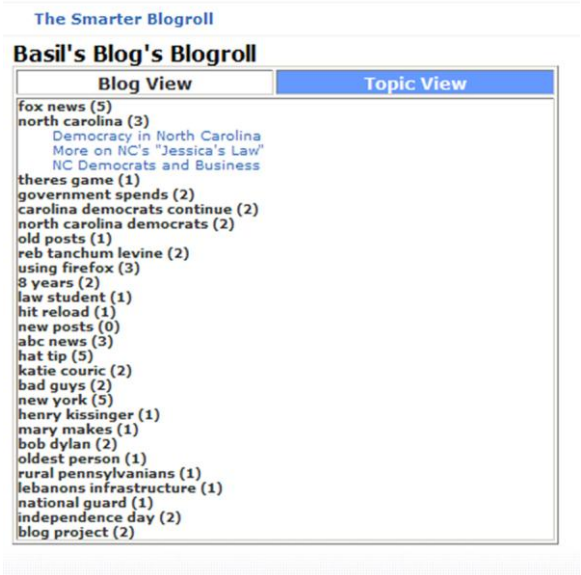

Fig. 2. The Smarter Blogroll, in topic view.

expanded simultaneously, and the posts for any topic can be hidden again after viewing.

In Figure 2, for example, we see that Basil's Blogroll has a strong interest in "fox news" and "north carolina." Clicking on "north carolina," in turn, shows three blog posts, which the user may follow through. By exposing topics through these two views, the tool facilitates the task of social browsing by allowing a user to quickly skim salient topics on any single blog or within their entire blogroll, i.e., trusted sources within their community.

Dataflow

The engine behind Smarter Blogroll is implemented as a server-side application. When a user configures a new account on the Smarter Blogroll, they provide as input a traditional blogroll, or list of URLs. The server follows each of these links to the blogs, then searches the pages for their RSS feeds and stores these as subscriptions. Periodically, the server downloads each of the RSS feeds to which it is subscribed, archiving a database of previous posts and their links. It then mines this database for topics.

When a user loads a Smarter Blogroll sidebar, the sidebar requests the top topics for that blogroll, and client-side code displays the results using the UI described above.

One of the major benefits of this design is that the central server can store both the blogrolls and the contents of the blogs. Users do not need to host any code of their own or add any mechanisms to their current blog. Indeed, as many blogrolls today are implemented as Javascript, it is a small effort to change from current blogrolls to a Smarter Blogroll. If the blogs are already in the database, the client can instantly populate the display; otherwise, the client must wait until the server has had a chance to download the relevant RSS feeds.

Topic Extraction

This paper is not intended to focus on the application of a specific topic extraction algorithm to the content of blogs. Indeed, a wide variety of topic extraction (e.g., [1, 11]) and summarization (e.g., [2]) techniques for non-technical language have been demonstrated in the research literature, many of which would be applicable to this project. Some past research has applied topic analysis to email corpora [14]. There has also been some

related work on opinion extraction [7] and summarization [16] for blogs. Based an upcoming 2006 TREC track that focuses on blogs [http://trec.nist.gov/tracks.html], as well as the emphasis on computational methods for blog analysis at the upcoming 2007 International Conference on Weblogs and Social Media [http://www.icwsm.org], we expect a flowering of research into topic understanding on blogs.

This paper illustrates that combining topic extraction with social search tools can provide a rich experience of navigation in blogs. Certainly, different back-end topic extraction algorithms could be applied to the Smarter Blogroll. This section describes the particular topic extraction method used to explore merger of topic extraction and social search.

In the current implementation, the Smarter Blogroll uses a tf-idf weighted listing of dominant bigrams and trigrams that occur in a blog post. In general, n-grams are sets of n words that appear consecutively in a document. For example, "appear consecutively" is a bigram in the previous sentence. We use a generic (and hand-tuned) stopword list to filter out common words out of the bigrams and trigrams. For example, we would not consider the trigram "that appear consecutively" because "that" is a stopword. In order to calculate tf-idf scores, posts are decomposed into a "bag-of-bigrams" and "bag-of-trigrams" representation, where the database stores the bigrams and trigrams for each post, along with the bigram or trigram's frequency count within that post. These bigrams and trigrams will be interpreted as the topics of the post.

The server generates a tf-idf score for all the potential topics in the posts on that user's blogroll: the term-frequency (tf) is taken to be the frequency of a given bigram against all other bigrams within the blog entry; the document-frequency (df) is taken the frequency of that bigram across all blog entries within that user's blogroll. tf and df are calculated for separately for bigrams and trigrams. We divide the term frequency by the logarithm of the document frequency to compute the tf-idf score. Thus, the tf-idf score is given by Eq. 1,

$$tf \cdot idf(x, p) = \frac{\frac{n_x}{\sum_k n_k}}{\log\left(\frac{|P|}{|\{p_i \mid x \in p_i\}|}\right)} \qquad \text{Eq. 1}$$

where x is a given term, p is a given post, n_x is the number of times x occurs in p, k ranges over all terms in p, |P| is the number of posts from the users' blogroll, and $|\{p_i \mid x \in p_i\}|$ is number of posts that contain x.

We scope the 'document frequency' to the entire blogroll because our focus is on the ideas that may be developing within this particular community. Other choices, including scoping to either the entire blog or to all blogs in the dataset, would give a sense for either what this blog or all blogs are discussing, respectively.

A list of top topics is then generated for each blog by taking the sum of the tf-idf scores for each topic that appears in a post on that blog, as given by Eq. 2,

$$sum(x) = \sum_p tf \cdot idf\,(x, p) \qquad \text{Eq. 2}$$

where x is a given term, p ranges over all posts on the blog in question, and *tf-idf*(x, p) is the tf-idf score of the term x in post p. The topics with the five highest scores for each blog are then displayed beneath that blog in the blog view. A list of top posts about these topics is also generated, based on the tf-idf score for the topic phrase in that post. Thus, each blog has a list of the top topics that appear on the blog, and each of those topics has a list of posts that are most about that topic.

Simultaneously, for the topic view, a similar list of top topics with associated posts is generated across all blogs. Again, Eq. 2 is used, but with p ranging over all posts in all blogs on the blogroll. To be listed under the topic view, a bigram or trigram is listed as a salient topic if the sum of its tf-idf scores across all documents in the blogroll is greater than 0.66. This value was chosen empirically to give a list of topics that was representative without being overwhelmingly long.

We tune these results slightly to match the special needs of our domain. First, we filter out bigrams that are substrings of high-frequency trigrams: for example, we trim "New York" and "York Times" if we have high-frequency trigrams for "New York Times." Second, we remove all results that do not occur in at least two entries across the blogroll, because tf-idf gives a very high score to a term that only appears once, which in this case is likely not a topic of interest to the community.

We chose to design this prototype around this relatively simple topic analysis method for several reasons. First, tf-idf scores can be computed in interactive time. This allows us to compute the weightings on-the-fly for each blogroll when it is requested. Second, these bigram and trigram-based methods generate easy-to-understand topics: the topic phrases come directly from the content of the blog posts they represent. In contrast, many clustering methods, even if potentially more accurate, generate harder-to-understand probability distributions of terms, to which it can be

Table 2. Statistics on the sizes of the blogrolls.

	Blogroll 1 Mean (*SD*)	Blogroll 2 Mean (*SD*)
Posts per Blog	7.0 (*2.9*)	21.8 (*11.6*)
Words per Blog	2383.3 (*1463.0*)	6270.8 (*3717.2*)
Words per Post	373.0 (*216.6*)	316.0 (*164.2*)

hard to apply a label. Because this is meant as a browsing technique, comprehensible labels are important to the design of the tool.

User Study

Methodology

Our user study is designed to test the proposition that the Smarter Blogroll can allow a user to understand the general activity on their blogroll. It is hard to operationalize browsing tasks, and so we looked for a task that would encourage users to try to get a "gist" of an entire blogroll, rather than having them search down a single needle in a haystack. We took the approach of encouraging users to look over many blogs on a blogroll at once to try to find general topics; they would then identify which topics were and were not discussed.

We collected two blogrolls on different topics from the "top 500" listing at blogrolling.com. For each blog on these blogrolls, we cached and stored locally the blog's front page, the posts linked to from those front pages, and the posts currently in the RSS feeds for each of those blogs. This fixed set allowed us to ensure that all users saw the same topics and results, no matter when or where we ran the study. In an effort to keep the task difficult but not impossible, we reduced both blogrolls to 25 entries. The first blogroll contained a total of 168 posts and 57200 words, and the second blogroll totaled 523 posts and 150,500 words. For further description of the sizes of the blogrolls, please see Table 2.

We then generated a collection of the topics of each blogroll. The authors, acting as independent raters, read all blogs on the blogroll and wrote down a list of topics that occurred in the blogs. The raters attempted to find topics that were general to more than one entry, but sufficiently specific to not encompass the entire blog. Topics were often paraphrases: thus, after seeing one blog discuss quilting and another discuss knitting,

"arts and crafts" was chosen as a topic. These topics were subjective, and created without reference to the tool. Ten items from each blogroll common to the two lists were selected as targets; a list of 20 irrelevant topics was also generated and added for each blogroll, leading to a list of 30 topics. We intended the task of identifying the relevant topics to be fairly challenging: in particular, we wanted enough targets and a short enough time that users would choose a browsing strategy, reading all blogs to understand their gists, rather than choosing a search strategy, looking for examples of posts about each topic.

We recruited twelve users with substantial blog-reading experience from a "bloggers" mailing list randomly assigning them to one of two groups and one of two orderings in a 2x2 study design. We used a within-subjects design, with users trying the Smarter Blogroll on one sample and a traditional blogroll on the other. Users in the first group saw "Blogroll A" with the Smarter Blogroll, while users in the second group saw "Blogroll B" with the Smarter Blogroll. Users in the first ordering used the Smarter Blogroll before they used a standard blogroll, while users in the second ordering had the reverse.

Users began their study with a brief warm-up task of reading a current, live blogroll through the Smarter Blogroll in order to get them accustomed to the interface and organization, as well as to iron out any technical problems. Subjects were then asked to take up to ten minutes to read the blogs on the blogroll, and were instructed to analyze the major topics covered on the blogroll. While reading, subjects were presented with the list of 30 topics during their reading, and asked to indicate which topics did or did not occur on the topics in the blogroll.

Users were rated on the number of correct hits and the number of false positives; in addition, they filled out a post-survey to report their experience.

Results

Our pool of subjects was very experienced with blogs: all except one either currently or had once read fifty or more blogs; all but one either actively maintained or had once maintained a blog of their own; half of those users had blogs with either blogrolls or lists of links.

Only half of the users read their blogs through RSS aggregators; there was no relation between using an RSS reader and reading many blogs. This suggests that some of our subjects might actively run into the challenges of keeping up with their blogs.

Users found the Smarter Blogroll much easier than the Classic; however, the Smarter Blogroll was not more effective for the task. After completing the study task using both the Smarter Blogroll and a classic blogroll, when asked to "rate the difficulty of the task with the Classic Blogroll" and "with the Smarter Blogroll" on a Likert scale from 1 (easy) to 5 (hard), users rated the task using the Smarter Blogroll as both much easier ($2.73 < 4.27$; $p = 0.001$; $n=11$; within-subjects paired t-test) and more enjoyable ($2.70 < 3.80$; $p = 0.017$; $n=10$; within-subjects paired t-test). Also, when whether "the topics displayed by the Smarter Blogroll gave a general sense of what the blogs are currently talking about" on a Likert scale from 1 (strongly agree) to 5 (strongly disagree), those participants who saw the Smarter Blogroll first said that the topic phrases provided by the tool gave a good general sense of what the blogs were talking about ($1.20 < 2.83$; $p = 0.028$; $n=11$; between-subjects paired t-test). Thus, the tool generally improved the user's experience of reading a blogroll.

Based on verbal feedback during the study, our subjects found the task to be very hard. They felt a substantial amount of pressure to search through the Smarter Blogroll looking for specific topics (often using the browser's 'find' functionality to search for specific keywords), and were disappointed that the Smarter Blogroll's bigrams did not align with the target topics.

Ultimately, the task may have been too difficult. On average, users correctly identified only 9.5 out of 20 targets and produced 6.3 false positives with no significant differences between the two blogrolls: neither blogroll made the task easier or more difficult. Furthermore, the Smarter Blogroll did no better than the classic blogroll: users using the Smarter Blogroll had neither a better accuracy nor fewer false positives then users with the usual form.

Discussion

The goal of the Smarter Blogroll was to aid users in getting a general impression of the current topics being discussed on a list of blogs. However, our results show that there was no significant difference between the number of topics correctly identified by users with the Smarter Blogroll and users with a classic blogroll. There are a number of reasons this may have occurred.

Examining the Experiment

Our study took one approach to testing a browse task, but our results demonstrate that it poorly reflected the actual task of browsing, instead driving users to search for information. Many users started at the top of the potential topics list and simply searched through the blogs looking for posts about that topic. Thus, our task produced a floor effect, in which all participants scored fairly poorly. We intended to produce a difficult task, so that participants felt a sense of urgency in completing it within the allotted time. However, as noted above, participants' performance indicates that the task may have been too difficult.

A different approach might ask users to write a summary or list of topics after finishing reading the blogs. However, such a list would be difficult to score, for reasons that are familiar in topic analysis research: is "the middle east" the same topic as "Lebanon"? In addition, it would be difficult to guarantee that users had browsed many blogs (rather than merely producing topics from the first two or three).

Another possibility is that tf-idf weighting on bigrams and trigrams may not be an appropriate method of topic extraction for blogs. We saw a number of topics where a term occurred in only two or three blog entries, giving it a very high tf-idf, even though it might not be considered as a "top topic" for the blog post. We are experimenting with other weightings for future work.

Last, topic extraction suffers from some of the same difficulties that other machine learning techniques do: if it is not accurate, users may find it inscrutable. Users clearly wanted some sort of additional information in their blogrolls, and seemed excited about the possibility of the improved blogroll. Yet our topics were not the metadata that they needed most.

This might help explain the significant order effect on whether the topics gave a good sense of what the blogs were about: users who saw the Smarter Blogroll first were more likely to report that it was a good summary then users who saw it second. When users started with the Smarter Blogroll, they felt limited (and some complained) when they had to use the classic blogroll without topic annotations. Even though the extracted topics did not help them perform better on the task, they felt that helped orient them within the blogroll. On the other hand, when users started with the classic blogroll, they were likely looking forward to the aid that the Smarter Blogroll would provide in completing their difficult task. When the Smarter Blogroll did not actually help that much in the task, they became disappointed in the Smarter Blogroll and rated it lower. We suggest that users are eager for the additional metadata, even if it was not helpful for the task.

Design Directions

The main contribution of these results is the finding that *metadata can aid in the facilitation of social browse tasks*. In future work, we will try to understand what sorts of metadata might better support user tasks.

There are a variety of places that we might look for additional metadata about blogs and blog entries. Blogs often have subtitles that may be more descriptive that the blog's name. For example, the blog Daily Kos derives its name from the last syllable of the organizer's first name Markos, which is not very indicative of the blog's content. However, the blog's subtitle, State of the Nation, makes in clearer that the blog deals with American politics and related issues. Also, individual entries carry their own titles, which can be much clearer than the title of the blog as a whole. To be sure, both of these are limited: the blog subtitle may turn out to be "a weblog that I wrote", and the title of an entry may well turn out to be "another thing". It was precisely to avoid these pitfalls that we originally designed the Smarter Blogroll. During the study, however, our users were disappointed by not having blog titles: they mentioned that even these scant cues were highly comprehensible to them.

Since blog entries are posted at a specific date and time, blogrolls could give an indication of post recency or frequency. Many blogs have explicit schedules, while others can acquire a certain rhythm to their posting patterns. Displaying or visualizing a blog's posting history may also give readers an impression of the blog's general rhythm, as well as how a current post fits into previous posting patterns.

In addition, blog entries are often annotated with categories and tags, which might act as a useful categorization and summarization tool. One of the main advantages to using tags as a means of summarization is that, because many bloggers tag their own posts, it does not require any extra computational overhead to automatically create categories or assign posts to those categories. Another advantage of tags is the possibility for bottom-up organization. No central authority creates standards dictating how tags are to be used. Since a tag can be any string of text, bloggers themselves determine what are appropriate tags and which tags to use for a given post. However, tags also have disadvantages. For one, not all bloggers use tags, and so a system that requires tags would not be applicable to all blogs. Also, due to their bottom-up organizational nature, the same tag may have very different meanings in different contexts. For example, the tag "public education" might have very different connotations on a political blog as compared to a parenting blog. Furthermore, tags do not inherently provide for hierarchical organization; I cannot create a tag "Venice 2006" that is a more specific version of the tag "Travel." While

tags are not the perfect form of categorization and summarization, they may prove to be an important piece of metadata to make available through blogrolls.

Beyond these forms of metadata that are already part of a blog, computed metadata provides additional possibilities. The tool described in this paper uses one specific topic extraction algorithm based on tf-idf scores for bigrams and trigrams, but other text mining approaches could be explored, such as PLSI and clustering [14], opinion extraction [7], or summarization [12]. Care must be taken, though, for these computed metadata can easily hide just as much information as they reveal. For example, a topic extraction algorithm carries the implication that its list of topics is definitive and all-inclusive, which masks the fact that there might be other topics not shown by the extraction algorithm. Computed metadata should be presented not as a definitive characterization of the blogs to which they are applied but rather as one possible view of those blogs.

Conclusion

This paper has explored combining social search and topic extraction through a prototype tool, the Smarter Blogroll, which allows a user to better understand what a list of blogs is about. Such a tool can facilitate some of the important functional and social purposes blogrolls serve, such as referring readers to related material, making a statement of interest, and enabling users to get a sense of what his or her blogging community is currently discussing. While the Smarter Blogroll did not significantly increase users' performance on the given study task, the results of our user study do serve to demonstrate some implications for the design of similar tools, as well as important possible future directions for research.

Even though their performance on the task did not improve, users greatly preferred having the blogroll augmented by a list of topics for each blog. This suggests that exploring other sorts of metadata about blogs could be potentially useful. While tf-idf weighted bigrams and trigrams were not beneficial for the study task, other topic extraction algorithms might help users.

The critical line of future work is to examine the reading behavior of blog readers, especially as it relates to blogrolls. During the course of our study, our users provided an almost constant stream of feedback about the ways that they go about reading blogs, their strategies for dealing with a long list of blogs, how they skim a blog to find relevant topics quickly, and other blog reading practices. Readers are an important part of blogging

communities, and understanding the ways in which they both contribute and keep up is crucial to understanding such communities. If we are to design tools to facilitate readers in their activity of reading blogs, we should first come to a greater understanding of the current practices surrounding blog reading.

Ultimately, the best evaluation of this sort of tool would come not from a laboratory study but rather from an *in-situ* study of the tool in actual use by bloggers. In such a setting, participants would be able to use the tool with their own blogroll and their own blogs, and determine whether such a tool aids in social browsing tasks over sources with which they are already familiar. Such a study should also focus on readers of blogs, examining how readers use classic blogrolls and react to tools such as the Smarter Blogroll. Hopefully, this can lead to the development of intelligent blogging tools that are equally useful for both authors and readers of blogs.

A design challenge highlighted in this study is to balance information content and legibility. The goal of the Smarter Blogroll is to enable users to quickly and easily browse through a blogroll. As noted by a number of our participants, since blogrolls are already rather long, annotating them with a list of topics for each blog makes the blogroll too large and unwieldy. Future research should explore other approaches to the user interface, such as expanding to show topics or other metadata on mouse over, fisheye magnification, topic clouds and other visualizations, or other forms of dynamic display.

Acknowledgements

Thanks to all our subjects for their participation in the study, to Liz Lawley for comments that guided the early design stages of this project, and to the members of MSR's Community Technology Group for their support and suggestions.

References

1. Bun, K.K. and Ishizuka, M., Topic extraction from news achive using TF * PDF algorithm. in Web Information Systems Engineering (WISE 2002), (2002).
2. Carenini, G., Ng, R.T. and Pauls, A., Interactive Multimedia Summaries of Evaluative Text. in 11th International Conference on Intelligent User Interfaces, (Sydney, Australia, 2006), ACM Press, 124-131.

3. CiteULike. A free service to organize your academic papers. http://www.citeulike.org.
4. Dennis, B.M. and Jarrett, A.C., NusEye: Visualizing Network Structure to Support Navigation of Aggregated Content. in 38th Hawai'i International Conference on System Sciences, (Big Island, HI, 2005), IEEE Computer Society.
5. Herring, S.C., Scheidt, L.A., Bonus, S. and Wright, E., Bridging the gap: A genre analyssi of weblogs. in 37th Hawai'i International Conference on System Sciences (HICSS-37), (Big Island, HI, 2004), IEEE Computer Society.
6. Herring, S.C., Kouper, I., Paolillo, J.C., Scheidt, L.A., Tyworth, M., Welsch, P., Wright, E. and Yu, N., Conversations in the blogosphere: An analysis "from the bottom up." in 38th Hawai'i International Conference on System Sciences, (Big Island, HI, 2005), IEEE Computer Society.
7. Ku, L.-W., Liang, Y.-T. and Chen, H.-H., Opinion Extraction, Summarization and Tracking in News and Blog Corpora. in Computational Approaches to Analyzing Weblogs: 2006 AAAI Spring Symposium (SS-06-03), (Stanford University, 2006), AAAI Press, 24-31.
8. Lenhart, A.B. Unstable Texts: An Ethnographic Look at How Bloggers and Their Audience Negotiate Self-Presentation, Authenticiy, and Norm Formation Graduate School of Arts and Sciences, Georgetown University, Washington, D.C., 2005.
9. Marlow, C., Audience, Structure, and Authority in the Weblog Community. in International Communications Association Conference, (New Orleans, LA, 2004). Retrieved September 2006 from http://www.researchmethods.org/ICA2004.pdf.
10. Nardi, B.A., Schiano, D.J. and Gumbrecht, M., Blogging as social activity, or, would you let 900 million people read your diary? in ACM Conference on Computer Supported Cooperative Work, (Chicago, Illinois, 2004), ACM Press, 222-231.
11. Ohtsuki, K., Matsutoka, T., Matsunaga, S. and Furui, S., Topic extraction with multiple topic-words in broadcast-news speech. in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98), (Seattle, WA, 1998), IEEE, 329-332.
12. Rau, L.F., Brandow, R. and Mitze, K., Domain-Independent Summarization of News. in Workshop on Summarizing Text for Intelligent Communication, (Dagstuhl, Germany, 1993), 71-75.
13. Schiano, D.J., Nardi, B.A., Gumbrecht, M. and Swartz, L., Blogging by the rest of us. in CHI Extended Abstraction on Human Factors in Computing Systems, (Vienna, Austria, 2004), ACM Press, 1143-1146.
14. Surendran, A., Platt, J.C. and Renshaw, E., Automatic Discovery of Personal Topics to Organize Email. in Conference on Email and Anti-Spam, (Stanford University, 2005).
15. Technorati. State of the Blogosphere, http://www.sifry.com/alerts/archives/000436.html, 2006.

16. Wu, Y. and Tseng, B.L., Important Weblog Identification and Hot Story Summarization. in Computational Approaches to Analysing Weblogs: AAAI 2006 Spring Symposium (SS-06-03), (Stanford University, 2006), AAAI Press, 221-227.