What Should Blog Search Look Like?

Marti A. Hearst*
School of Information
UC Berkeley
Berkeley, CA, 94720
hearst@ischool.berkeley.edu

Matthew Hurst
Microsoft Live Labs
Bellevue, WA, 98006
mhurst@microsoft.com

Susan T. Dumais
Microsoft Research
Redmond, WA, 98052
sdumais@microsoft.com

ABSTRACT

Blog search has not yet reached its full potential. In this position paper, we suggest that more could be done to accommodate the task of finding good blogs to read, especially with respect to matching a desired taste or style of writing. We propose a faceted navigation interface as a good starting point for blog and author search, and that search oriented around people and their writings will lend itself well to advanced interfaces. We also argue that blog search is probably best integrated with search on other forms of timely social media for the task of determining what is currently being thought about a particular topic.

Categories and Subject Descriptors

H.5 [**Information Interfaces and Presentation**]: User Interfaces; H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Human Factors, Design

1. INTRODUCTION

Why are there so few successful blog search engines, and why don't the existing ones make good use of the special properties of blog data? The blog search offered by Ask and Google are quite similar to web and news search, while Yahoo and MSN abstain altogether. Technorati, perhaps the best-known blog search engine, does innovate in several ways. It computes and shows an "authority" score for each blog, it organizes blog posts into topic categories, and it allows readers to browse popular posts both by their recency and by the amount of "attention" they have received.[1] However, much more could be done; Ramakrishnan & Tomkins [17] state that current interfaces for search over blogs and bulletin boards do not make good use of the special features of this kind of data.

Additionally, to date there has been very little academic work on search interfaces for blogs or other social media [2]. Most of the attention surrounding blogs has been focused on the link structure among people and posts, and on the study of the dynamics of information flow within these link networks (e.g., [4, 5, 9, 11]).

A good way to make choices for user interface design is to do needs assessments and other forms of user research. One excellent study of this kind has been done: Mishne & de Rijke [14] did an extensive query log analysis of a blog search engine. They found that 52% of 500 randomly chosen ad hoc queries contained named entities: people, products, companies, etc. Of the rest, 25% consisted of high-level topics such as "stock trading," "gay rights," etc., and the remaining 23% were navigational queries, adult queries, and other disparate types. Mishne & de Rijke determined that most of the named-entity queries were requests to learn what is being said currently about that entity, while the more general queries were often attempts to find blogs or posts on a topic of interest. They also found that for the most popular queries, 20% of these were related to recently-breaking news items. They also found a marked difference from web queries issued over the same time period. One can interpret the findings of Mishne & de Rijke to suggest that people are currently searching blogs for: 1) discussion about current events and people, and 2) thoughts on a topic generally.

In addition to information needs, blog search interface design must take into account how blogs differ from other kinds of online textual information that is currently searched. Mishne [13] notes that blogs exhibit differences in *language* (more informal), in *structure* (a dense micro-community topology, typed links such as blogrolls and trackbacks with distinct meanings), and the importance of *recency*. Other attributes that should be taken into account include the fact that documents are *no longer full HTML pages* (a blog post is an element in a feed or a subpart of an HTML page), the information is often *subjective* or opinion-oriented, and perhaps most importantly, the data is *people-centric*.

How do we translate these findings into how a search engine should present information? A perhaps radical point of view is: traditional web information is useful for archival and reference information, but social media is something quite

*Written while visiting Microsoft Research.

[1]The authority score is based on the number of blogs linking to a given blog within the last month. The attention score is a weighted rank based on time, number of links, rate of new links, Technorati Authority, and the Technorati Authority of linking blogs. http://support.technorati.com/support/siteguide/channels

SSM'08, October 30, 2008, Napa Valley, California, USA.

different and is best accessed in new ways. While standard search results are pages and answers to questions, social media results take the form of aggregate analyses and access to people. That said, some of the content produced by social media is valuable long after the initial posting and can benefit by being searched by standard means.

Thus, rather than one interface for blog search, we propose thinking about the issues in terms of three different kinds of search tasks:

1. Find out what are people thinking or feeling about X over time.
2. Find good blogs/authors to read.
3. Find useful information that was published in blogs sometime in the past.

These tasks are reflected to some degree in the TREC blog track [16]. Types 1 and 2 are also somewhat reflected by commercial blog search engines, but much can be done to improve those (for instance, Google only occassional returns a list of blogs to read, and this list is usually quite short). Type 3 is somewhat covered by web search engines that index older blog posts. Each of these ideas is discussed below in terms of what it means for the search user interface.

2. TASK TYPE 1: SEARCH THE PULSE OF THE POPULACE

Authors such as Battelle [1] see the exposure of timely user-generated content and opinions as a way of taking the mental pulse of the populace. This desire to understand the conversation about "what is happening now?" as reflected in blogs is, to some degree, what is typically supported in blog search today. (Most blog search engines do not allow search over posts that are more than a few months old.)

The developers of the TREC blog track recognized this important aspect of blog search by making the first task for the track be to identify which posts contained opinionated information, later extending this to recognizing the polarity of those opinions [16]. The lively, opinionated aspect of blogs and other forms of social media are being explored in applications other than blog search. Work by Mishne & Rijke [15] plotted statements about moods over time, based on LiveJournal mood tags, and sites like WeFeelFine.org extract out statements about peoples' emotional states and present them in beautiful visualizations. Research has been done (e.g., [4, 5]) on how to help market researchers use the results of sentiment mining – based on the substantial body of research that has already been done on this problem – to help get a timely understanding of reactions to products and policy proposals alike. The temporal/timeliness aspect of blogging plays an important role in this kind of analysis.

Applications that make use of opinion mining most likely need their own specialized interfaces, but there is nevertheless a role for general search for the "what is being said about X now?" task. Rather than a blog search exclusively, it is probably preferable for an opinion/rating/impression search engine to index numerous forms of social media, including time-sensitive blogs, micro-blogs (e.g., Twitter, FaceBook walls), product review sites, the analysis and opinion sections of traditional news media, and perhaps the letters to the editor and news analysis sections of academic journals and other publications.

How should the interface differ from what is being offered by blog search today? It should organize and aggregate the results better, and by having a focus on author information, including who has commented on the post, and who has blogged about the post. Blog post aggregators such as Sphere and Blogrunner were successful (and were acquired by other companies) because they show information about blog posts related to a well-defined stimulus such as a news article in mainstream media. Blogpulse provides a very simple version of this idea by plotting how frequently different words are mentioned over time across posts. But much more sophisticated processing and presentation could be done.

If we decide that rather than a separate blog search for "what is being said now," we should have a more general social media search with a short time horizon, then where does this leave blog search engines? Perhaps with a different question, namely: what is a good blog for me to read?

3. TASK TYPE 2: CHOOSE A BLOG OR AN AUTHOR TO READ

The information seeking modality of *serendipity* is supported today by following links within blog rolls, by web sites that stream popular or unusual stories (as done with "memetracker" web sites such as TechMeme and BuzzFeed), and by human-generated ratings sites such as Digg and StumbleUpon.

But there is currently no good, systematic way to find good blogs with particular properties. Ounis, Macdonald & Soboroff [16], in summarizing the two years of the TREC blog track, note that "[T]he prevalence of blog directories suggests that there is not yet a suitable way to identify blogs with a recurring interest in a topic area ... [and] remains an open problem." They set up what they call the *topic distillation* task as a mechanism for developing research advances in this area. Specifically, the task was to find blogs with a central and recurring interest in X where X is some topic, and X is the subject of the majority of the posts within the blog. They note that the techniques that worked well for topic distillation were markedly different than those that worked well for opinion finding.

Although the TREC task is a big leap forward, the formulation of the tasks is missing a number of dimensions that would be useful for improving blog search:

- Characteristics/Quality of blogs. The current TREC topic distillation task focuses (for good reason, it is a starting point) solely on relevance of blogs but does not take into account quality metrics, for example:
  - How many posts are original content versus commentary on other content?
  - How much of the relevant content for the topic of interest does the blog cover?
  - What is the style or tone of the writing? (Discussed in more detail below.)
- Subtopics within topics. A reader may want to find blogs that provide high-quality commentary on one topic specifically within a general subject area, for example, commentary on a particular television show or on a particular model of motorcycle. Often these are interspersed with high-quality commentary on other

related topics, such as other TV shows or other vehicles. A blog selection interface should allow for the automatic creation of a feed reader on only the subtopics of interest across several high-quality blogs simultaneously, with little or no additional work needed on the part of the user.

- Information relating to the people who read and write the blog.
  - Who are the people who do the interacting on the blog, including in comments?
  - Whom does the blog link to, and which others are linked to it? What forms of media link to it?
  - How many people write for this blog? What are their reputations?
  - How many people post comments for the authors of the blog? What is the quality of the comments?
  - Does this blog link to others with similar or different viewpoints?

Kritikopoulos et al. [10] incorporate some of these features to develop a better ranking algorithm by taking into account information about the link graph among blogs, including the number of links between blogs, the number of common tags or categories they share, the number of users who have posted comments in both blogs, and the number of times the two blogs link to the same news story. These factors are incorporated into a PageRank-like algorithm to successfully improve rank ordering of blog posts. This is an intriguing use of social and linking aspects of blog data, but is used to improve the standard interface.

3.1 A Faceted Blog Finding Interface

Probably the best interface for the blog finding problem, given the multiple dimensions to be satisfied, is a *faceted interface*, which has been shown to be highly usable for navigating information collections [6] and personal information [3], and is widely used on vertical web sites today.

Reflecting the discussion above, facets would include blog genre (group blog, journal-style, etc), blog style (informal, humorous, expert; see discussion below), frequency of update, proportion of original content, and detailed topic categories. Facets should also include detailed information about people, including authors, author affiliation, sites at which the author tends to post, the political leanings or expertise of the author, and so on. They should also show information about readers, including who comments on the blog, who reads the blog, etc. There are some interface challenges with showing this kind of information as there is a long tail of readers and linkers for any given person, but central tendencies can be usefully shown for browsing, with search for specific individual tightly integrated into the interface (as is done in standard faceted interfaces).

3.2 More on People Search

People should be a central focus of social search. Ramakrishnan & Tomkins [17] discuss the idea of *content claiming*, which is the capability to aggregate all content authored by one person into a single place, whether that content be formal publications, blog posts, or comments and reviews provided on a variety of web sites. Mishne [13] also suggests that the research community focus on developing profiles of individual bloggers.

In addition the faceted navigation attributes discussed above, blog search should allow users to explicitly look for people who have written on a topic, and see their writings aggregated across different media boundaries, see who they link to and who links to them, what their social networks are, and see other cues that may reveal their expertise and experience. The web site Spock has the beginnings of a highly-linked people search; this kind of interface could be usefully integrated within blog search. The web site MarkMail is experimenting with brushing-and-linking interfaces for exploring author information integrated with posts to technical bulletin boards. More generally, the people-oriented aspect of blog search would probably lend itself well to more advanced interfaces including visualization.

3.3 Matching on Blog Style or Personality

Finding a good blog to read is somewhat akin to finding a good book to read. It follows from experience with book and film e-commerce sites that a dimension that might be of great interest for finding a good blog is that of the *style, taste, personality* or *attitude* of the author, thus allowing the user to indicate which of the following they would like to see in a blog's writing style: witty, snarky, serious, empathetic, silly, scientific, business-oriented, artsy, new-age-y, religious, etc. This is to be distinguished from, but related to, topic, which might include discussion of policy, industry, science, current events, and so on. These in turn differ from genre, which includes personal diary, corporate blog, group blog, etc. (Of course, blogs are a genre onto themselves.)

Standard Text Classification A straightforward approach is to run a classifier over blogs, or authors within blogs, or individual posts, that characterize the attitude along the dimensions named (snarky, artsy, serious, etc). These choices, along with the other important dimensions, would be one facet in a faceted interface as discussed above. Kale et al. [9] show the power of using links typed by opinion polarity information (trusting or not trusting) to determine link-minded blogs, and features like these that make use of social and opinion data may be useful for such a classifier.

Relevance Feedback + Collaborative Filtering However, for subjective responses such as those corresponding to attributes best characterized as taste, it has been shown that relevance feedback and collaborative filtering based methods tend to work well [7, 12]. It might be best to, after eliciting the subject matter that the user is interested in, show them sample posts from the various styles, and allow them to indicate which they do and do not like, and then either show blogs that have been determined via a categorization system to be similar in style, or use collaborative filtering to recommend blogs that people who have taste similar to yours have liked in the past.

Implicit Selection The suggestions above assume the user is in the mood to put some work into selecting a blog (which we argue is not an unrealistic assumption in some cases). But alternatives that make use of implicit information, as opposed to explicit specification, should also be considered. Something as simple as a system that keeps track of which blogs a user ends up reading posts from, whether found from search results, or from following links or email pointers, could be a good source of recommendations. Although attempts to infer what web pages to view based on

those already seen are not particularly successful, blog preferences gathered in this manner should be more successful, because it is somewhat similar to book recommendations. Recommendations gathered in this way would be useful for finding blogs on topics of interest generally, as well as for building a model of what kind of taste the reader prefers.

Descriptive Queries Another alternative is to encourage *descriptive queries* and attempt to return posts from blogs that match along other dimensions in addition to content. Recent work suggests that longer queries are more descriptive of the explicit information need, and that the appropriate type of response can sometimes be inferred from the structure of the query [8]. For example, the query "how to choose a divorce lawyer" is asking for advice and would benefit from a serious discussion. By contrast, "how to cope with divorce" might benefit from a site written from an empathetic viewpoint, while "help me laugh about my divorce" might be looking for a humorous take on relationship issues.

3.4 Other Approaches

An entirely different approach to blog selection is proposed by Leskovec et al. [11] where the goal is to view the minimum number of blogs that cover all of the topics that the user wants to be sure not to miss.

4. TASK TYPE 3: SEARCH ARCHIVED BLOG POSTS

Especially in the early days, blog posts were seen as ephemeral and disposable; what matters is only what is said now. However, with the increasing professionalism of blogs, as well as the increased participation of posters, it is the case that many posts would be valuable to look at even long after the events in question have passed. A useful tool would be a kind of blog "trash compactor," reorganizing a blog after the fact (especially if it was to be mothballed) by grouping articles on related topics together, featuring those that were most original and best written, and perhaps grouping with other archives on the same topic. Of course, at the point that a blog post becomes mature and useful, it can be treated much the same as venerable web pages are by web search engines. It should also be noted that incorporating linking information to good blog posts into standard web search results should improve those results.

5. CONCLUSIONS

Blog data is social, temporal, and more generally, multifaceted. For the problem of selecting a blog to read, we propose a faceted interface which highlights different attributes of interest, with a focus on people and on matching the taste preferences of the reader. For the task of "taking the pulse of the blogosphere," we suggest that blog data be integrated with other social media and that the existing work on tracking trends and aggregating views is heading in the right direction.

6. REFERENCES

[1] J. Battelle. *The search: how Google and its rivals rewrote the rules of business and transformed our culture*. New York: Portfolio, 2005.

[2] J. Cho and A. Tomkins. Guest Editors' Introduction: Social Media and Search. *IEEE Internet Computing*, 11(6), 2007.

[3] E. Cutrell, D. Robbins, S. Dumais, and R. Sarin. Fast, flexible filtering with Phlat – Personal search and organization made easy. *Proc. CHI 2006*, 2006.

[4] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. *Conference on Knowledge Discovery in Data*, 2005.

[5] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. *Conference on Knowledge Discovery in Data*, 2005.

[6] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, and K. Yee. Finding the flow in web site search. *CACM*, 45(9), 2002.

[7] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1), 2004.

[8] M. Kaisser, M. Hearst, and J. Lowe. Improving search results quality by customizing summary lengths. *Proceedings of ACL/HLT'08*, 2008.

[9] A. Kale, A. Karandikar, A. Java, T. Finin, and A. Joshi. Modeling trust and influence on blogosphere using link polarity. *Proceedings of ICWSM'07*, 2007.

[10] A. Kritikopoulos, M. Sideri, and I. Varlamis. BlogRank: ranking weblogs based on connectivity and similarity features. *Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, 2006.

[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. *Proceedings of SIGKDD'07*, 2007.

[12] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1), 2003.

[13] G. Mishne. Information access challenges in the blogspace. *Proceedings of IIIA*, 2006.

[14] G. Mishne and M. de Rijke. A study of blog search. *Proceedings of ECIR*, 2006.

[15] G. Mishne and M. de Rijke. Capturing global mood levels using blog posts. *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[16] I. Ounis, C. Macdonald, and I. Soboroff. On the TREC Blog Track. *Proceedings of AAAI*, 2008.

[17] R. Ramakrishnan and A. Tomkins. Toward a PeopleWeb. *COMPUTER*, 2007.