The Role of the Author in Topical Blogs

Scott Carter
EECS Department
University of California, Berkeley
Berkeley, CA 94720
sacarter@cs.berkeley.edu

ABSTRACT

Web logs, or blogs, challenge the notion of authorship. Seemingly, rather than a model in which the author's writings are themselves a contribution, the blog author weaves a tapestry of links, quotations, and references amongst generated content. In this paper, I present a study of the role of the author plays in the construction of topical blogs, in particular focusing on how blog authors make decisions about what to post and how they judge the quality of posts. To this end, I analyzed the blogs and blogging habits of eight participants using a quantitative analysis tool that I developed, a diary study, and interviews with each participant. Results suggest that authors of topical blogs often do not create new content but strive to, often follow journalistic conventions, use the content of their blogs as a reference tool for other work practices, and are connected as a community by a set of source documents. Results also show that Instant Messaging is useful as an interview medium when questions center around online content.

Author Keywords

Blogs, authorship, content analysis, qualitative methods

ACM Classification Keywords

H5.m. Information interfaces and presentation (*e.g.*, HCI): Miscellaneous.

INTRODUCTION

> *[The author's] only power is to mix writings, to counter the ones with the others, in such a way as never to rest on any one of them.*
> Roland Barthes [1]

> *The author vision blinds us to the importance of the commons — to the importance of the raw material from which information products are constructed.*
> James Boyle [2]

In their respective works, Barthes and Boyle called for the end to the single author regime, arguing instead that the attribution of content should be distributed to the larger community out of which the producer arose. Indeed, blogs seem to fit well with this approach, especially Barthes description of the work of the author as one who "mix[es] writings." But not included in Barthes' and Boyle's arguments is the perception of the author themselves. Do they consider themselves stewards of a commons or are they motivated also by less altruistic issues? The contribution of this work is to provide insight onto the notion of authorship with respect to blogs. I address this by looking at both the practice of blog authorship as well as the ways by which blog authors judge the success of posts.

One difficulty with studying blogs is that they span a wide spectrum of uses. While prior work by Schiano *et al.* and Nardi *et al.* investigated the habits of personal bloggers, in this study I concentrate on blogs that focus on a particular topic (*e.g.*, *filter* or topical blogs) run by one person [3–5]. To further constrain the work, I selected blogs based in the United States that concentrate on intellectual property issues. The restriction of the study to a particular community limits the generalizability of findings but facilitates analysis of social trends such as topics of mutual interest and trusted sources.

In this paper, I describe the method used to study the practice of authoring blogs, a three-pronged approach using quantitative analysis of blog posts, a diary study, and interviews. I then report and discuss the findings: that authors of topical blogs often do not create new content but strive to, often follow journalistic conventions, use the content of their blogs as a reference tool for other work practices, and are connected as a community by a set of source documents, and that Instant Messaging (IM) is useful as an interview medium when questions center around online content.

METHOD

I used a snowball sampling method to recruit eight people (six men and two women; four students, three law professionals, and one technology professional) who maintain blogs. Each blogger identified their blog as relevant to the intellectual property community. I used a three pronged approach to understand person's blog posting habits and decisions. First, I built a link and quote analysis tool and ran it on each blog. Second, I developed a Web-based diary study questionnaire that each blogger responded to over the course of two weeks. Third, I interviewed each blogger about their general blogging habits and decision making processes as well as about the data gathered from the analysis tool and diary study. I chose this combination of methods to be able to compare and contrast actual posted content with authors' perception of the practice.

I ran an analysis tool on one year's worth of postings for each blog, save one whose author started blogging only nine

CHI 2005, April 2-7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

Table 1. Post analysis

	Avg	Min	Max
Avg word length	318.0	164.0	643.9
Avg external links	4.7	2.1	7.7
Avg words per external link	66.9	48.5	111.0
Quotes per post	0.6	0.2	1.7
Avg quote word length	105.5	71.5	166.3
Percent of post quoted words	23.9	9.0	39.0

Table 2. Frequency of link sources, normalized

Type	Normalized frequency
Topical personal blogs	37.5
Topical filters or group blogs	35.0
Other and unattributed	11.1
Popular news	8.4
Law sources	8.0

months prior. The analysis tool I developed crawls a given blog and extracts post, hyperlink, and quote data (as few blogs included pictures, image analysis was excluded). The tool uses blog-specific hypertext markers to extract data. It uses *div* tags or comments to locate postings; *a* tags to locate links, and *blockquote* tags to locate quotes. Also, the tool discovers the source of quotes by analyzing local text within the post prior to the quote. Specifically, if there was only one link prior to the quote in the blog text that link was presumed to be the source of the quote. If there were several or no links in prior text, then an algorithm was applied that weighted links in prior and post text using two metrics: nearness and the structure of the sentence in which the link appeared (*e.g.*, links embedded in prepositional phrases such as "...from the [link]" were rated higher than others). If no links were discovered in the text, or source confidence was below a pre-set threshold, the source for that quote was not attributed.

Participants were instructed to fill out the Web-based questionnaire I developed every time they posted to their blog over a two week period. The purpose of the diary was to gain more insight on posting decisions than is revealed in posts themselves. The questionnaire consisted of questions regarding the content of the post, sources used in the post, and resources the author used to discover sources.

I interviewed each participant after running the analysis tool on their blog and gathering diary study data. The interviews were roughly divided into two parts: the first set of questions attempted to discern how participants found information to post and how they judged the quality of posts, while the second set of questions were clarifications of data collected by the analysis tool and diary study. I conducted two interviews in person, two via phone, and three via IM. While I tried to maintain consistency between the different mediums as much as possible, I also noted pitfalls and advantages of each medium, concentrating on IM as its use is not often reported.

LINK AND QUOTE RESULTS

As Table 1 shows, compared to the findings of Herring *et al.* of a general set of blogs, the blogs in this study had slightly longer posts and many more external links per post [3].

Table 3. Frequency of quote sources, normalized

Type	Normalized frequency
Popular news	34.4
Topical filters or group blogs	22.5
Law sources	15.2
Other and unattributed	14.7
Topical personal blogs	13.2

Table 4. Specific links that appear in at least half of the blogs

Type	Frequency
Documents	8
Documents (on group blog)	5
Documents (on personal blog)	1
Documents (on law code site)	1
Documents (on personal site)	1
Personal blog site	7
Topical sites	3
Group blog post	1
Personal blog post	0

Thus, the blogs I analyzed had a much higher link density (or fewer words per link). We also found much higher rates of quoted words. While Herring found that quotes accounted for roughly 10% of post matter, quoted words accounted for roughly 27% of the post matter in the blogs I analyzed.

Analysis of the normalized frequency of domain references showed differences in the reliance on sources for links versus quotes (see Tables 2 and 3). A normalized frequency analysis reveals the sources that a group of bloggers rely on without overemphasizing sources used by more profuse bloggers. Normalized frequency was calculated by first normalizing the frequency of link references and quote references to a domain within each blog and then summing the normalized frequencies across all blogs. For links, participants in our study rely upon topical news filters and group blogs as well as other personal blogs that are also topical. While topical filters or group blogs are also sources for quotes, popular news articles are quoted more often than other sources.

Table 4 shows a categorization of specific links that appeared in at least half of the blogs surveyed. This data provides insight regarding the themes that run through this community. As the table shows, the common references that defined this community of bloggers were a set of documents as well as links to other blogger's sites.

DIARY AND INTERVIEW RESULTS

Because the participants blog at different and fluctuating rates, responses to the diary study varied considerably. Over the two week diary, participants completed a questionnaire a median of 3.5 times, a maximum of 21 times and a minimum of zero times. The primary source for posts was the Web (33 posts), followed by documents (5 posts), discussions (two posts) and others (four posts). The most frequent reason participants listed for their discovering the source was via an e-mail message, but overall participants did not often respond to this question.

Each participant was interviewed with an emphasis on understanding their model of a good post and the ways that they judge the quality of a post. Another area of interest was the practice of searching for topics for a post and how that practice interacted with other work practice.

Post quality

Participants overwhelmingly commented that a good post is one that contributes new information or, to a lesser extent, extensive commentary about some issue on which the participant is an expert. Some participants included the timeliness of the post with respect to its subject material as being important as well. When asked to specify a particular post that they had written that they judged to be high quality, respondents usually chose posts that had much lower link and quote densities than average for their blog.

Amongst posts in which the participant did not provide much new content (*e.g.*, link-based posts), participants said that it was best to link to completely new information or at least source information, bypassing other filters and news sources. Four participants reported archiving a document on their own blog that either had not been published digitally before or was likely not to be available on the Web indefinitely. In fact, in two cases participants reported that such a document was the most heavily trafficked link on their site.

Six participants reported linking to source documents whenever possible. For this community, "source documents" almost always meant court decisions or specific laws.

Participants reported judging the quality of a post primarily by trackbacks (links from other blogs to their post) or by their own analysis of server traffic. Another metric that most participants used was links from blogs with a much larger perceived audience than their own. Participants did not attribute much value to the size and quality of comments left on the blog.

Finally, six participants said that they followed journalistic convention when updating posts — explicitly marking changes and using extra text or color to call attention to the fact that the post had been changed. The two participants who did not follow this convention instead followed a self-described "wiki" model, updating posts frequently and undeclaredly.

Blog as reference archive

Six participants described the use of their blog as an archive tool directly linked to their work practice. In these cases, posts often served as way to save information that would later be used in the construction of other documents. These participants also reported using other topically related blogs in the same manner. One participant commented that, "I use [another participant's] blog before Google: its easy to find the information I need and there's usually a good analysis in the post." Also, two participants commented that they thought their work limited what they might blog either because they were worried about the ramifications of others at their work reading the blog or because a topic was so commonly discussed at their work place that it did not seem worth reiterating on their blog, even if it might be a novel topic outside of their workplace.

Audience

Seven participants reported struggling with writing their posts in such a way that both a topic savvy audience as well as a lay audience could get something out of their blog. Participants said that their goal was to make their posts as broadly understandable as possible, but that usually time constraints restricted them from doing so. Also, while most participant's blog posts consistently addressed intellectual property in some way, three participants specifically commented that they often interjected more personal posts so that they were not "taken too seriously" by their readers.

Finding issues to blog

Participants relied on news feeds and e-mail lists to find sources for their posts. All participants also reported perusing topic related blogs and news sites. While most participants perused content throughout the day, two reported making this a primary task for one to two hours a day.

Interviewing using IM

I followed recommendations of Voida *et al.* with regards to interviewing over IM [6]. On one hand, I found that overall the interview took longer (on average 82 minutes versus 63 for phone and face-to-face interviews). On the other hand, I found it much easier to follow references to links that came up in the interview. Specifically, interviewees embedded 10, three, and 16 links into the IM window. Of those links, six, two, and seven, respectively, were references to specific blog posts with long URLs. While references to specific blog posts came up in face-to-face and phone interviews, it was much more difficult to navigate to them: often the interview would say "search for [topic]" to direct me to a post, but I rarely located the correct post.

Another positive of using IM included being able to peruse source links and other information on the Web while the interview was taking place. In addition, participants reported that they were able to get some other work done or at least take breaks during the interview without it significantly impacting the results. However, it was sometimes unclear when a participant had completed a thought. One participant used a previously adopted practice of using ellipses to indicate continuations to handle this problem (*e.g.*, "I don't know, let me look at that post... Oh, that came from an aggregator...").

DISCUSSION

The results presented above suggest that authors of topical blogs often do not create new content but strive to, often follow journalistic conventions, use the content of their blogs as a reference tool for other work practices, and are connected as a community by a set of source documents. The results also show that IM is useful as an interview medium when questions center around online content.

The composition and quality of posts

Topical blog authors are as Dadaists selecting readymades who yet strive to be Bauhaus craftspeople and do occasionally succeed. That is, they strive to create original content and commentary but are also content to spread a found meme that they consider important or interesting. As mentioned above, most posts had a much higher number of links per post than do blogs at large, suggesting that they more than anything else the authors refer and comment on issues raised

elsewhere. Data regarding the composition of these links shows that many of these memes originate from other topical or group blogs. This practice seems to be motivated both by an altruistic desire to inform as well as a desire to spread the author's name. The former motivation led to the authors' writing style dilemma: while they want to spread a piece of information as rapidly as possible, they also want to write their posts to cover as broad an audience as possible. The later motivation is tightly coupled to this in that the author's name is tightly coupled to any content in the blog post. The authors expect that other authors would refer to the place where they discovered information originally. Authors exercised this expectation through the monitoring of server data and trackbacks.

Similarly, while blog authors reported that they would rather refer to source information, quotes from popular news organizations and topical filters constitute a significant amount of their post content. This may simply be due to the fact that original, source materials (*e.g.*, court decisions and laws) are few in comparison to the documents that surround them (*e.g.*, articles and columns). However, as commented that one of the most important contributions that they can make is to select important source information and distill it for a lay audience, it may be that they are in a similar mind set when they quote from the popular news: selecting important information and processing it for others to better understand.

Also, authors' post-hoc judgement of the quality of a post is tightly linked to whether or not their post is discovered by popular personal or group blogs. While many participants did not report this directly, it can be inferred from the fact that most participants commented that raw traffic is one the most important metrics that they use to judge the success of a post, and a post usually only becomes heavily trafficked when it reaches popular blogs. Even blog authors who insisted that their blog was for them alone still analyzed server logs or regularly tested their links for hits in major search engines. Ultimately, then, reputation is directly linked to perceived quality.

Blog author models

Topical blog authors are often described using newspaper metaphors, usually as "columnists" or "pundits" [3]. While these terms describe some of their habits well, such as their approach to updating posts, many authors would rather be perceived as a role more similar to that of the beat reporter, breaking new stories, documenting a unique experience, or uncovering new material about a particular topic.

In addition to reporting to the public about issues of perceived importance and building a reputation for the author, the blog post also serves as an archive and reference tool. Many bloggers either work in an area related to their blog or are involved in other activities similar to the theme of their blog. As such, they often use blog posts as a source of reference when crafting other documents, such as school papers or legal briefs. New blogging software or tools must be careful to support these parallel functions.

The blogging community

As reported above, the common references that defined this community of bloggers were a set of documents as well as links to other blogger's sites. The documents were largely the type of source material that the authors agreed was the most important to link to: specific court decisions and codified laws. Thus, while the prevalent blogging habit may have been to link and reference other blog postings and popular media, there was agreement in the community of the relative importance of "original" documents. The other common set of links, those of other bloggers, were generally those that the authors discussed as being influential and the ones from which they aspired to be linked. Thus, this blogging community is one that is highly aware of its hierarchy and is tightly linked to a certain type of document.

Interviewing using IM

I found that the benefits of using IM for interviews outweighed the drawbacks. However, many of the benefits derived from the fact that the interviews centered around digital documents, making it much easier both to refer to specific posts and to search for information in a blog that might influence the specific questions asked in the interview. Also, participants were unsure how to communicate that they had completed a thought. A simple two-state button on an IM interface might alleviate this problem considerably.

CONCLUSIONS AND FUTURE WORK

While authors of topical blogs have an interest in the commons and share ideas and findings with others, they seek respect and expect to be acknowledged by others who benefit from their efforts. Also, they take the authorship of documents into consideration when linking and quoting in their posts, often relying on mainstream sources but valuing content produced by those involved in the actions they report.

Any applications of these findings to blogging tools and techniques is difficult because the use of the blog format is rapidly evolving. But future work might concentrate on ways for blog authors to express their perceived quality of a post or at least provide mechanisms to categorize and annotate content based on the intended effect of a post not only its subject matter. Also, future studies could compare results across different communities.

ACKNOWLEDGEMENTS

I thank P. Duguid, G. Nunberg, and J. Mankoff.

REFERENCES

1. R. Barthes. *Image-Music-Text*. Hill and Wang, 1978.
2. J. Boyle. *Shamans, Software, and Spleens: Law and the Construction of the Information Society*. Harvard University Press, 1996.
3. S. Herring, L. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *Proceedings of HICSS*, 2004.
4. B. A. Nardi, D. J. Schiano, and M. Gumbrecht. Blogging as social activity, or, would you let 900 million people read your diary? In *Proceedings of CSCW*, pages 222–231, 2004.
5. D. J. Schiano, B. A. Nardi, M. Gumbrecht, and L. Swartz. Blogging by the rest of us. In *Extended abstracts of CHI*, pages 1143–1146, 2004.
6. A. Voida, E. D. Mynatt, T. Erickson, and W. A. Kellogg. Interviewing over instant messaging. In *Proceedings of CHI*, pages 1344–1347, 2004.