Leave a Reply: An Analysis of Weblog Comments

Gilad Mishne
Informatics Institute, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam
The Netherlands
gilad@science.uva.nl

Natalie Glance
Nielsen BuzzMetrics
5001 Baum Blvd, Pittsburgh, PA 15213
natalie.glance@buzzmetrics.com

ABSTRACT

Access to weblogs, both through commercial services and in academic studies, is usually limited to the content of the weblog posts. This overlooks an important aspect distinguishing weblogs from other web pages: the ability of weblog readers to respond to posts directly, by posting comments. In this paper we present a large-scale study of weblog comments and their relation to the posts. Using a sizable corpus of comments, we estimate the overall volume of comments in the blogosphere; analyze the relation between the weblog popularity and commenting patterns in it; and measure the contribution of comment content to various aspects of weblog access.

1. INTRODUCTION

Weblog comments serve as "a simple and effective way for webloggers to interact with their readership" [9]; they are one of the defining set of weblog characteristics [21], and most bloggers identify comment feedback as an important motivation for their writing [19, 4]. Despite this, comments are largely ignored in current studies of large amounts of weblog data, typically because extracting and processing their content is somewhat more complex than extracting the content of the posts themselves.

In this paper, we present the first large-scale study of comments in the blogspace; the research questions we address are:

- What is the number of weblog comments and the volume of comment content? How does it compare to the volume of weblog posts?
- To what extent does usage of comments improve access to weblogs in typical tasks, such as weblog search?
- What relation exists between the amount of comments on a particular weblog or post and its popularity, as measured by traditional prestige metrics?
- What knowledge can be mined from comments and the discussions taking place within them?

The rest of this paper is organized as follows. In the remaining part of this section we briefly survey related studies. We follow with a description of our corpus and how it was obtained. Next, in Section 3, we analyze the contribution of the comment contents to weblog access tasks. In Section 4 we study the association between commenting patterns and some indicators of weblog popularity, such as amount of incoming links. Section 5 takes a closer look at one particular aspect of comments, namely, the type of discussion carried out in them. Our conclusions appear in Section 6.

1.1 Related Work

As mentioned earlier, quantitative studies of weblogs focus on post data, leaving out the comments. A single exception to this is the work of Herring *et. al* [5], which examine a random sample of 203 weblogs. In this sample, a relatively small amount of comments is found (average of 0.3 comments per post); however, this sample is too small for definite conclusions regarding the entire blogosphere, and the comments themselves are not further analyzed. Qualitative analysis of weblogs, on the other hand, sometimes refers to comments explicitly. Both Trevino and Gumbrecht study the importance of weblog comments to the "blogging experience" [19, 4], reaching similar conclusions: comments are regarded by most bloggers as vital to the interactive nature of weblogs. Krishnamurthy studies the posting patterns to a specific weblog following the September 11 events, finding that insightful posts attract the largest number of comments [7]. De Moor and Efimova [1] discuss weblog comments in a larger context of weblog conversations; among their findings is user frustration about the fragmentation of discussions between various weblog posts and associated comments, indicating that for users the comments are an inherent part of the weblog text, and they wish to access them as such. Comments can be considered implicit links between people. Extracting comments enriches the model of the social network of bloggers and readers, and may be used for enhancing studies of weblog communities and the interactions between them, including the burstiness work by Kumar *et. al* [8] and the in-depth analysis of a specific community done by Wei [20]. Finally, weblog comments are a source of search engine optimization spam; this is discussed in [11].

2. DATASET

In this section we describe the comment corpus we studied and how it was built.

2.1 Comment Extraction

The vast majority of weblogs support various types of syndication – standards which enable easy access to updated web content by machines. Syndication facilitates applications which access contents of weblogs directly (e.g., RSS

WWW2006, May 22–26, 2006, Edinburgh, UK.
.

readers), bypassing the task of parsing HTML and mining content from it.

Unfortunately, with the exception of a small number of weblog software vendors and blogging hosts, comment information is currently largely unsyndicated. A preliminary study we conducted prior to collecting our comment corpus indicated that less than 2% of comment content is currently available in syndicated form. We expect this figure to increase, as blogging platforms develop and new standards allowing syndication of comments (e.g., RSS 2) are adopted. However, creating a comment extraction mechanism from weblog HTML content was required to account for existing weblog comments. We implemented a relatively simple wrapper for weblog pages; in a nutshell, the extraction performed by it includes the following stages:

- Identify the "comment region": the continuous section within the HTML page most likely to contain comments. Typically, this is between the end of the post and a page footer (or sidebar).
- Inside this region, identify lists of dates which have characteristics typical of comments.
- Expand each date to a complete comment by analyzing the text around it.

The process is somewhat similar to the extraction of weblog posts from HTML content described in [2]. We leave out the details of the wrapper's implementation as this is not the main focus of this paper.

Coverage. To test the coverage of our wrapper, we manually evaluated its output on a set of 500 randomly-selected weblog posts from our corpus; in this set, 146 posts (29%) contained comments. Coverage was tested by comparing the manual comment extraction and the automated one, measuring the percentage of posts for which extraction was correct, as well as the percentage of posts with no comments which were correctly identified as such.[1] The results of this evaluation are given in Table 1.

Set	Correct	Incorrect
Posts with no comments	342 (97%)	12 (3%)
Posts with comments	95 (65%)	51 (35%)

Table 1: Comment Extraction Evaluation

Note that for 11 out of the 51 comment extraction failures (21% of failures), the number of comments and their dates were correctly extracted, but the content was not. This means that for analyses which do not take content into account (such as determining the average number of comments per post), the wrapper's accuracy is over 70%. In addition, 23 out of the 51 failures—almost half—originated from non-English pages; our coverage on English pages only is close to 80%.

2.2 A Comment Corpus

We collected a set of approximately 645,000 comments posted to weblogs between July 11th and July 30th, 2005. The set was obtained using the following steps:

1. Collect all weblog posts in the Blogpulse [3] index from the given period containing a permalink.
2. Remove "inactive" weblogs – weblogs which had low posting volume in the months preceding the analyzed period.
3. Fetch the HTML of the remaining permalinks, and run the extraction process described earlier.

In each of these steps, some content is missed: in the first stage, posts with no permalinks are ignored. The next stage filters a large amount of single-post-only blogs (which account for a significant percentage of total weblogs [14]), as well as a lot of spam weblogs – an increasingly popular phenomenon [18]. In the final stage, sources of missing content are multiple: broken links, hosts which restrict crawling the post HTML (e.g., LiveJournal), and the wrapper's incomplete coverage. Overall, based on estimation of the amount of content missed in every stage, we believe that our comment corpus includes more than one quarter of all comments posted to weblogs, in the entire blogspace, during the 20-day period for which data was collected.

Table 2 contains some statistics about the collection.

Weblog posts	685976
Commented weblog posts	101769 (15%)
Weblogs	36044
Commented weblogs	10132 (28%)
Extracted comments	645042
Mean comments per post	0.9
Mean number of days in which commets were posted, per post	2.1
Comments per post, excluding uncommented posts	
Mean	6.3
StdDev	20.5
Median	2
Comment Length (words)	
Mean	63
StdDev	93
Median	31
Total corpus size	
Words	40.6M
Text	225MB

Table 2: Corpus Statistics

As expected, the number of comments per post follows a power-law distribution, with a small number of posts containing a high number of comments, and a long tail of posts with few comments; a plot of the number of weblogs and posts having a given number of comments is shown on a log-log scale in Figure 1, with the best-fit power-law coefficient. The distribution of the comment lengths is similar – few long comments, and much more shorter ones.

Total Comment Volume. Based on our corpus and the estimates regarding the coverage of the comment extraction process, we estimate that the number of weblog comments in the entire blogosphere is comparable to the number of posts in active, non-spam weblogs: this means that the total number of comments is somewhere between 15% and 30% of the size of the blogosphere as reported by major weblog search

[1] Incorrect identification of comments on a post without comments usually occurred in highly-irregular weblogs, where the wrapper misinterpreted blogrolls and the post list.

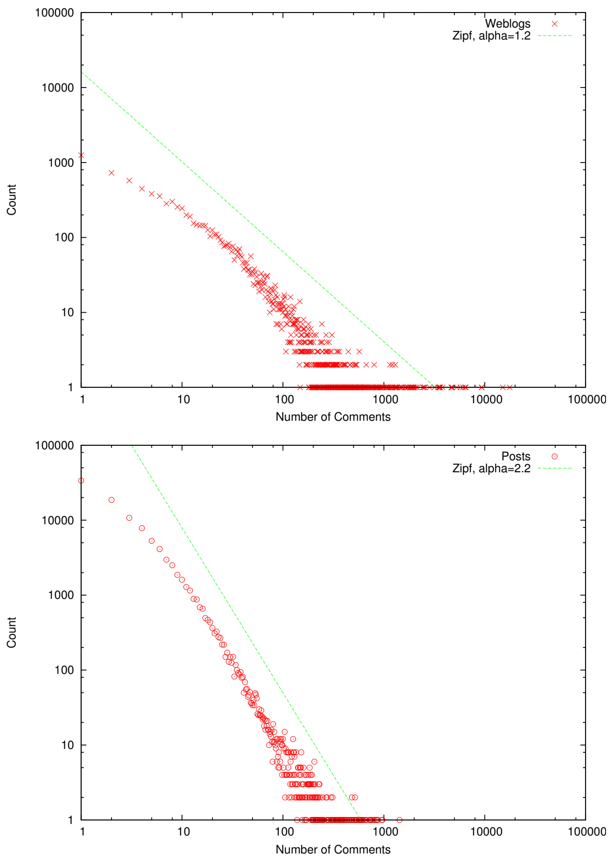

Figure 1: Power-law distribution of the amount of comments per weblog (top) and post (bottom).

engines such as Blogpulse. At the time of writing (February 2006), weblog posts are added at a rate of over 700,000 a day (including spam and inactive weblogs): assuming our estimates are correct, this means a daily comment volume in the order of 150,000 comments.

On average, comments are shorter than weblog posts (in terms of text length); comparing the average length of a comment to the average length of a post in the corpus we described, we estimate that the textual size of the "commentsphere" is 10% to 20% of the size of the blogosphere. Note, however, that influential weblogs tend to have more comments than non-influential ones (see Section 4); in some cases of top-ranked weblogs, the volume of comments far exceeds the volume of the posts themselves. By overlooking comments, much of the conversation around many influential blogs is being missed.

Comment Prevalence. An additional issue to consider when studying weblog comments is that some weblogs do not allow commenting at all. While the vast majority of blogging platforms support comments, bloggers themselves sometimes choose to disable this option, to prevent flaming, spam, and other unwanted effects; other bloggers permit comments but moderate them, by manually reviewing submitted comments before publishing them or allowing comments from trusted sources only. This, naturally, reduces the overall potential volume of the commentsphere. Reports on the amount of weblogs permitting comments are mixed; a low figure of 43% appears in the random sample examined in [5], while the community-related sample studied in [20] shows that more than 90% of the weblogs enabled comments (both studies do not report on the actual number of commented weblogs). In our collection, a random sample of 500 weblogs shows that over 80% of weblogs allow users to add comments to the posts, but only 28% of weblogs actually had comments posted.[2] The increase in comment prevalence, compared to [5], can be attributed to the development of blogging software in the 2.5-year period between the two studies.

2.3 Links in Comments

Overall, our comment corpus contained slightly more than 1 million HTTP links, an average of 1.6 links per comment. This number includes "signature links" – links which the comment author leaves as identification, in many cases linking back to her weblog. For the same time period, weblog posts themselves contained close to 20 million links. A examination of the top-linked-to domains in comments, in comparison to the top-linked-to domains in posts, shows similar results: the top domains are weblog communities such as `blogger.com` and `xanga.com`, data sharing websites such as `flickr.com` and `groups.yahoo.com`, news sites, and large retailers such as `amazon.com`. We found no substantial differences between the linking patterns in comments and in posts, and do not expect comments to contribute significantly to algorithms involving link analysis of weblogs.

Having said that, in some weblog domains (e.g., LiveJournal) there is very little direct linking from post to post, and social behavior is centered instead around commenting. Thus, following the commenting behavior in these domains is crucial for understanding the social network and identify communities. In such domains, commenting can be mapped to linking – after which link-analysis methods used in link-rich weblogs can be applied.

3. COMMENTS AS MISSING CONTENT

Following our initial analysis of the amount and volume of comments, we turn to evaluate to what degree the absence of weblog comments affects real-life weblog access. The natural task to turn to is weblog search – retrieving blog contents in response to a specific request from a user. In this section, we study the usage of comment content in this context.

We collected a set of 40 queries submitted to the weblog search engine at Blogpulse.com during the 20-day period for which the comment corpus was extracted. For each of these 20 days, we randomly selected two queries from the most popular 5 queries submitted by Blogpulse users during that day.[3] Example queries from this set are "space shuttle" (July 14th), "Clay Aiken" (July 16th), and "Mumbai floods" (July 29th). We then retrieved the results of each query from two separate indices: an index of all comments we extracted as detailed earlier, and a subset of the regular Blogpulse index containing all weblog posts from the corresponding period. While our comment index contained 645000 comments, the weblog post index contained over over 8M posts (this number includes spam and inactive weblogs,

[2] Both of these figures are likely to increase if including LiveJournal weblogs, which are often commented.

[3] By "most popular", we mean queries which were submitted by the largest amount of different users.

making it higher than the total number of posts shown in table 2).

Recall. First, we study the contribution of comment content to the recall of the retrieval – the number of (relevant) returned results. While recall is typically not an important measure of web retrieval, this is not the case for weblog search. In fact, comparisons of weblog search engines focus on recall (as well as presentation issues such as duplicate detection), e.g., Hodder's comparison in [10]. This is because weblog searchers tend to view results sorted first by recency, then by relevance – in other words, they are more interested in complete coverage over the recent hours or days than in web-style relevance estimations.

To measure the improvement in recall, we compared the list of permalinks retrieved by searching the post index to the list of permalinks retrieved from the comment index (multiple comments from the same post permalink were considered as a single hit for that permalink). For each query, we analyzed the overlap between the lists, as well as the contribution of each source separately. For example, for the query "space shuttle", a total of 7646 permalinks were retrieved from both indices; of these, 7482 (96.9%) were retrieved from the post index only, 164 (2.2%) were retrieved from the comment index only, and 74 (0.9%) were retrieved from both.

	Posts Only	**Comments Only**	**Both**
Mean	93.2%	6.4%	0.5%
StdDev	9.1%	8.7%	0.7%
Median	96.9%	2.6%	0.2%
Minimum	64.3%	0%	0 %
Maximum	100%	33.3%	2.4%

Table 3: Recall Contribution of Comments

Table 3 shows the aggregated results over all 40 queries, using the same percentage view as used in the example. Keeping in mind that our corpus is estimated to contain around a quarter of all comments posted during the period (whereas our post corpus is more or less complete), we see a notable contribution of content in comments to the overall recall. Extrapolating our observations to account for the comments which are not in our corpus as a result of the extraction process (see Section 2), we estimate an addition of 10%–20% "hits" for a query on average, given a complete index of all weblog comments; the median addition would be lower at 5%–15%, due to a small number of queries with very high contributions from comment contents (in our experiments, these included both queries with many hits such as "rss" and queries with few ones, such as "Tighe"). In particular, it is interesting to note the relatively small overlap between the results of the comment search and the post search – suggesting that comments often add new terms to the contents of the post, terms which assist in retrieving it given a query.[4] Also worth noting is the high standard deviation of the contribution of comments to recall, indicating that, for some queries, comment content is vital to complete coverage of the blogosphere.

Precision. As outlined earlier, in the weblog domain precision is not typically used to compare retrieval results. Most weblog search engines present their results sorted by date, assuming that recent results are of higher importance to the searcher; typically, results from the same date are sorted according to some static ranking of weblogs, based on an estimation of the weblog's popularity. Examining the top results produced by this type of ranking with Blogpulse.com for the queries in our test set, we experienced good overall retrieval quality: the majority of top-ranked posts were indeed relevant for the queries.

However, weblog searchers are possibly interested in more than the topical relevance usually used to evaluate retrieval. Analyzing a community weblog, Krishnamurthy [7] observes that "The number of comments per post is perhaps the truest and most diagnostic metric of the nature of communication on a weblog. The posts that are most insightful or controversial get the most comments. Those that are pedestrian do not get many comments". This led us to believe that while topical precision itself is not greatly modified by using comments, they do provide access to a different perspective of weblog posts, namely, the impact on their readers.

Evaluation of this new precision angle is complicated; in fact, its definition by itself is beyond the scope of this paper. To support our claim anecdotically, we experimented with a method for reranking the top 100 results produced by the "standard" ranking method according to the number of the comments associated with the weblog posts. We tested this method on 10 different queries, examining the top-10 ranked results and assessing them for relevance. We found that this method, while preserving the same early precision levels as the "standard" ranking method, produces top-ranked results which are more discussion-oriented and attract more feedback from users. However, our evaluation was cursory and more rigorous tests need to take place to support this.

To summarize, while usage of comments does not alter the precision numbers, it offers users a different scheme for viewing results. This is comparable with the two sorting criteria weblog search engines currently offer users, date and relevance: the precision *value* doesn't change, but its *meaning* does. What is meant by the change in meaning of precision? When results are sorted by date, precision is the proportion of timely *and* relevant posts; when results are sorted by relevance, precision is (typically) the proportion of popular/authoritative *and* relevant posts. Adding comment text to the index changes again the flavor of precision: the metric now becomes the proportion of highly discussed relevant posts with relevant comments.

4. COMMENTS AND POPULARITY

Cursory examination of weblogging patterns, as well as intuition, suggests that a large number of comments is consistent with the influence level a weblog or post has - the degree to which it is read, cited, linked to, and so on. In this section we attempt to substantiate this assumption using our corpus.

To measure weblog popularity we use two indicators: the number of incoming links as reported by the Blogpulse index, and the number of page views for weblogs that use a

[4] In several cases, we observed almost-empty posts, containing just a link to an article or another web page with a short remark such as "unbelievable"; the comments to the post contained actual content and keywords, supplying the context to the post and enabling its retrieval.

public visit counter such as Sitemeter.[5] In total, there were 8824 weblogs for which we had both page view counts and inlink information [17]; of these, we found comments in 724 weblogs.

Tables 4 and 5 compare the number of incoming links and page views for weblogs with no comments and blogs with varying levels of comments.

Number of comments	Count	Average page views	Average incoming links
0	8104	453.7	66.7
> 0	724	812.9 (+79%)	267.1 (+300%)
Breakdown:			
1–10	186	423.2 (-7%)	130.4 (+95%)
11–50	260	485.3 (+7%)	158.5 (+137%)
51–100	115	650.8 (+43%)	261.2 (+291%)
101+	163	1894.6 (+317%)	600.3 (+800%)

Table 4: Weblog popularity, compared to the number of comments

Average comment length (words)	Count	Average page views	Average incoming links
0	8104	453.7	66.7
> 0	724	812.9 (+79%)	267.1 (+300%)
Breakdown:			
1–10	46	782.4 (+72%)	327.7 (+391%)
11–50	291	388.3 (-14%)	156.6 (+136%)
51–100	260	978.5 (+116%)	309.1 (+363%)
101+	127	1457.8 (+221%)	412.2 (+518%)

Table 5: Weblog popularity, compared to the average size of comments

Clearly, commented weblogs are substantially more read and linked to. However, there is a chicken-and-egg situation here: assuming a fixed percentage of weblog readers post comments, weblogs which have more incoming links and more readers are more likely to have higher amounts of comments. Nevertheless, the existence of many comments in a weblog post is clearly an indication for popularity of the post, and unlike other measures (such as indegree count) does not require analysis of the entire blogosphere.

4.1 Outliers

While we witnessed a good correlation between the level of comments and the weblog popularity on average, we also encountered various exceptions: high-ranking weblogs with no or little comments, low-ranking weblogs with many comments, and so on. We now discuss some of these cases.

"Too few" comments in high-ranked weblogs. Many weblogs, particularly high-ranked ones, impose some moderation on reader comments, or disable them altogether; this is typically done to prevent spam and other forms of abuse. Of the top-10 ranked weblogs with no or few comments we checked, all employed some sort of comment moderation, leading us to believe that these outliers are mostly artificial.

"Too many" comments in low-ranked weblogs. Most weblogs that appeared to have substantially more comments than expected given their viewership and incoming link information turned out to be weblogs of the personal-journal flavor, where a relatively small group of the blogger's friends used the comment mechanism as a forum to converse and interact. Many of these comments did not relate directly to the post, and resembled a chat session more than other comments in our comments.

An additional class of weblogs which have a high number of comments, given their link indegree, consisted of weblogs that are popular with the non-techy crowd, such as fashion or celebrity weblogs – presumably, readers of these weblogs tend to use links less than the more technologically-oriented readers (or, alternatively, do not blog at all).

Highly-commented posts in a given weblog. Some posts in our corpus have a very large number of comments, compared to the median in that weblog: some examples appear in Table 6. In general, it seems such posts are either related to highly-controversial topics (usually, politics), or posts which were cited in mainstream media or in other sources directing a high level of traffic towards them.

- http://blog.qiken.org/archives/2005/07/harry_potter_th.html
Review and spoilers from the book *Harry Potter and the Half-Blood Prince*, posted close to its release date.

- http://www.riehlworldview.com/carnivorous_conservative/2005/07/natalee_hollowa_44.html
Deals with a search for a missing person, and a reward offered for information about her.

- http://www.majorityreportradio.com/weblog/archives/002569.php
Part of a discussion about U.S. policies in Iraq.

- http://www.geenstijl.nl/mt/archieven/006443.html
Dutch political weblog, post deals with Dutch policies towards Muslim immigration.

- http://www.rickey.org/blog/2005/07/constantine_mar_6.html
About the TV show "American Idol" and its participants.

- http://www.rosie.com/2005/07/06/wednesday-2/
Personal weblog.

- http://www.thinkprogress.org/2005/07/21/breaking-bloomberg-reporting-that-rove-
  - libby-may-be-subject-to-perjury-charges
Political weblog, post deals with a U.S. political scandal.

- http://hurryupharry.bloghouse.net/archives/2005/07/12/thinking_aloud.php
Personal commentary about reactions to terror attacks in London.

Table 6: Examples of highly-commented posts, compared to other posts from the same weblog

5. DISCUSSIONS IN COMMENTS

Weblog comments provide a rare opportunity to explore user responses to online content. Excluding weblogs, wikis, and message boards, feedback on web sites is typically submitted through forms and email, and is not available publicly. A small number of personalized websites have guestbooks—a leftover from earlier internet days—but even those are used to provide feedback about the entire site, rather than about a particular topic or section. In contrast, weblogs which allow commenting allow direct, personal, mostly unmoderated discussion of any post in the weblog. In this section we explore one aspect of these discussions, namely,

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

controversy in comment discussions.

Examining our comment collection we identified various types of comments – among those are personal-oriented ones (posted by friends), comments thanking the author for raising an interesting issue or pointing to additional related content, and so on. One class of comments we found particularly interesting was the set of *disputative* comments, comments which disagree with the blogger (or with other commenters), forming an online debate. We hypothesized that these comments can be used to identify controversial topics, authors, newspaper articles, and so on. An example of two comment threads from the same weblog, one of them disputative, appears in Table 7. In this section we set to identify this type of comments computationally.

5.1 Detecting Disputes in Comments

We address the task of finding comment threads indicating a controversy as a text classification problem. We trained a decision tree boosted with AdaBoost[6] using a set of 500 manually annotated comment threads (the examples in Table 7 are taken from this set). In total, 79 (16%) threads in the set were labeled "disputative".

We follow with a description of the features we used for our classifier.

Feature Set

- **Frequency counts** - the basic and most popular feature set used in text classification tasks [16]. We used counts of words and word bigrams in the comments, as well as counts of a manually constructed small list of longer phrases typicallly used in debates ("I don't think that","you are wrong", and so on).
- **Level of Subjectivity.** With a large amount of training data, the frequency counts would have captured most important words and phrases distinguishing controversy from other discussions. However, given our limited training data, we chose to measure the *level of subjectivity* of the comments separately from the frequency counts. By this we are referring to usage of phrases such as "I believe that" and "In my opinion", which tend to appear in disputative comments more than in other comments. To capture these types of phrases, we compared the language used in the encyclopedia entries of Wikipedia[7] to the language used in the user discussions about these entries. This meant building a language model for the encyclopedia entries themselves (2GB of text), a model for the discussions (500MB of text), and comparing them using log-likelihood, a standard corpus divergence metric [6]. The top phrases found using this comparison include "I don't", "you have to" and so on; the subjectivity level we assigned to a comment thread consisted of the sum of log-likelihood values of the phrases occurring in it.
- **Length Features.** Observing that disputative comments tend to be longer and appear in longer threads, we added features for the average sentence length, the average comment length in the thread, and the number of comments in the thread.

`mainstreambaptist.blogspot.com/2005/07/neo-con-plan-to-help-military.html`

Post:
The Neo-Con Plan to Help the Military
The New York Times published an article yesterday, "All Quiet on the Home Front, and Some Soldiers are Asking Why," that has a paragraph revealing the neo-conservative's plan to assist the military fulfill its mission in Iraq. Here it is . . .
It will be interesting to see how the bankers and lawyers and doctors and engineers in Oklahoma respond to this call for support. If they can't sell their program here, they can't sell it anywhere.

Comments:

1. It's about time all those that voted for the warmonger in charge to put up or shut up.

2. Bruce, this is exactly what my son, Spc. ccsykes, was talking about when he made the following comment on your blogpost - "Iraq Imploding" - "Lack of support from the people of the United States, low morale in the Military and our policies have already lost this war."

3. Marty, you are right and so is ccsykes.

4. One of the more shocking moments, I thought, was when Bush counseled us to go out and shop in response to the ramping up of terrorism. Though I want us out of Iraq as soon as possible, I think we owe Iraq the . . .

5. ditto

`mainstreambaptist.blogspot.com/2005/07/reclaiming-americas-real-religious.html`

Post:
Reclaiming America's Real Religious History
Kudos to Marci Hamilton at AlterNet for her outstanding article on "The Wages of Intolerance." She does an outstanding job of reclaiming America's real religious history from the revisionists who want to make . . .

Comments:

1. I think that the author's candor about American history actually undermines her argument. First, she . . ....

2. Anon, it is obvious to me that you don't know Bruce personally. He speaks the truth. Perhaps one day the scales will fall off your eyes as well . . .

3. Perhaps Bruce could be more persuasive in proving his admittedly controversial (and I would say wild) assertions.

4. I've given a little thought to something I wrote to Bruce earlier: "You yourself seem to be guilty of wanting to impose . . ." It would be absolutely futile for me to attempt to engage in a reasonable discussion there; it is just as futile to do the same here.

5. I've watched with great interest as this blog has evolved from a discussion of Mainstream Baptist concerns and demoninational issues into a hyper-political, left-wing campaign against . . .

6. you can always tell that someone does good work, because someone is going to get angry about it. all this man is doing is standing up to what he believes are historic baptist principles.

7. mt1, I suggest that you read the description of the blog that is the top of each page. I also sign my full, real name to everything I write.

8. Anonymous, commenting on your comments has been very theraputic for me. God bless you and good luck on your new . . .

Table 7: Disagreement in comments: non-disputed post (top) and thread including disagreement (bottom), from the same weblog

[6] We experimented with other types of classifiers such as Winnow, with similar but slightly lower results.

[7] http://en.wikipedia.org

- **Punctuation.** We used both frequency counts of the various punctuation symbols in the text, and special features indicating usage of excessive punctuation (this has been shown to be effective for certain tasks of text classification, e.g., [15]).

- **Polarity.** The sentiment analysis method described in [13] was used to identify the orientation of the text of the comments. The intuition here is that disputes are more likely to have a negative tone than other types of discussion.

- **Referral.** While studying our corpus, we noticed that comments which disagree with the weblog author (or with another commenter) are likely to contain certain types of references to previous content or authors. Typical references are a quote (from the weblog post or from another comment), referring to previous authors by name, and increased usage of second-person form. We implemented simple heuristics to detect such referrals and used their existence—as well as information regarding how early they appear in the comment—as additional features; for example, a direct quote as the first sentence of the comment was taken to be a referral.

5.2 Evaluation

Using a 10-fold cross validation on our manually annotated corpus, we obtained an accuracy level of 0.88, as shown in Table 8. As this is an unbalanced distribution, comparison to a baseline is difficult (see, e.g., [12]) – a maximum-likelihood classifier would have achieved an overall F-score of 0.84 by classifying all threads as non-disputative, but would have little meaning as a baseline as it would have yielded an F-score of 0 on the disputative comments only.

	Precision	Recall	F-Score
Non-disputative comments	0.92	0.96	0.94
Disputative comments	0.72	0.58	0.65
Overall	0.88	0.89	0.88

Table 8: Comment Extraction Evaluation

The following are the most important features utilized by the classifier, in decreasing order of importance:

- Existence of a referral quote in the first part of the comments
- Usage of question marks early on in comments
- Counts of phrases from the manually-built disagreement lexicon
- Number of comments in the thread
- Polarity of the first sentence of the comments
- Level of subjectivity

Among the words and word bigrams which were relatively important features are pronouns and negating words such as "not" and "but".

Using the classifier on the entire comment corpus resulted in about 21% of the comment threads to be tagged disputative, suggesting that comments are used, in many case, for argumentative discussion. Anecdotal examination of the disputed comment threads, in particular those assigned a high confidence by the classifier, suggests that these threads do contain a fair amount of controversial discussions. Table 9 contains the top terms appearing in disputed comment threads (excluding stopwords), showing what's "on the blogger's mind"; clearly, politics prevail as the central topic of debate.

Iraq	Government
Money	Country
America	Political
Bush	Women
Power	White House
Church	Media
President	School
United States	Children
Muslims	The Media
Supreme Court	The Constitution

Table 9: Disputed Topics

6. CONCLUSIONS

We presented a large-scale study of weblog comments, a domain often neglected in computational studies of weblogs; according to our analysis, comments constitute a substantial part of the blogosphere, accounting for up to 30% of the volume of weblog posts themselves.

In terms of the contribution of comment content to current weblog access, we show that usage of comments improves weblog retrieval (in terms of number of results), and is beneficial for ranking weblog posts in new, potentially useful ways.

We discuss comments as an indicator of the popularity of weblog posts and weblogs themselves, and find—as expected—that a wealth of comments in a weblog is a good indication for the significance of the weblog.

Finally, we offer a novel way to determine the level of controversy caused by a weblog post by analyzing the type of comments written in response.

Acknowledgements. The authors wish to thank Matthew Hurst, Kamal Nigam, Matthew Siegler, Robert Stockton and Takashi Tomokiyo for many useful discussions and for their assistance. Thanks also to our anonymous reviewers for useful suggestions.

7. REFERENCES

[1] A. de Moor and L. Efimova. An Argumentation Analysis of Weblog Conversations. In *The 9th International Working Conference on the Language-Action Perspective on Communication Modelling (LAP 2004)*, 2004.

[2] N. Glance. Indexing the blogosphere one post at a time. In *Third International Workshop on Web Document Analysis (WDA2005)*, 2005.

[3] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004, at WWW '04: the 13th international conference on World Wide Web*, 2004.

[4] M. Gumbrecht. Blogs as "protected space". In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004, at WWW*

'04: the 13th international conference on World Wide Web, 2004.

[5] S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. In *The 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*, 2004.

[6] A. Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37, 2001.

[7] S. Krishnamurthy. The multidimensionality of blog conversations: The virtual enactment of september 11. In *Internet Research 3.0*, 2002.

[8] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[9] C. Marlow. Audience, structure and authority in the weblog community. In *The 54th Annual Conference of the International Communication Association*, 2004.

[10] Mary Hodder. A Comparison of How Some Blog Aggregation and RSS Search Tools Work, 2005. napsterization.org/stories/archives/000500.html and napsterization.org/stories/archives/000502.html, accessed November 2005.

[11] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *First International Workshop on Adversarial Information Retrieval on the Web, at WWW '05: the 14th international conference on World Wide Web*, 2005.

[12] M. Monard and G. Batista. Learning with skewed class distributions. In *Advances in Logic, Artificial Intelligence and Robotics*, pages 173–180, 2002.

[13] K. Nigam and M. Hurst. Towards a robust metric of opinion. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.

[14] Perseus Development Corporation. The Blogging Iceberg - A Blog Survey, 2003. www.perseus.com/blogsurvey/iceberg.html, accessed November 2005.

[15] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk E-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.

[16] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[17] M. Siegler. Private communications, October 2005.

[18] Technorati, 2005. The Second Web Spam Summit, www.technorati.com/weblog/2005/09/47.html, accessed November 2005.

[19] E. M. Trevino. Blogger motivations: Power, pull, and positive feedback. In *Internet Research 6.0*, 2005.

[20] C. Wei. Formation of norms in a blog community. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff, and J. Reyman, editors, *Into the Blogosphere; Rhetoric, Community and Culture of Weblogs*, 2004.

[21] D. Winer. What makes a weblog a weblog?, 2003. blogs.law.harvard.edu/whatMakesAWeblogAWeblog, accessed November 2005.