New Metrics for Blog Mining

Brian Ulicny[a], Ken Baclawski[a], Amy Magnus[b]
[a]VIStology, Inc., 5 Mountainview Dr., Framingham, MA 01701
[b]Air Force Office of Scientific Research, 875 North Randolph Street, Suite 325 , Arlington, VA 22203

ABSTRACT

Blogs represent an important new arena for knowledge discovery in open source intelligence gathering. Bloggers are a vast network of human (and sometimes non-human) information sources monitoring important local and global events, and other blogs, for items of interest upon which they comment. Increasingly, issues erupt from the blog world and into the real world. In order to monitor blogging about important events, we must develop models and metrics that represent blogs correctly. The structure of blogs requires new techniques for evaluating such metrics as the relevance, specificity, credibility and timeliness of blog entries. Techniques that have been developed for standard information retrieval purposes (e.g. Google's PageRank) are suboptimal when applied to blogs because of their high degree of exophoricity, quotation, brevity, and rapidity of update. In this paper, we offer new metrics related for blog entry relevance, specificity, timeliness and credibility that we are implementing in a blog search and analysis tool for international blogs. This tools utilizes new blog-specific metrics and techniques for extracting the necessary information from blog entries automatically, using some shallow natural language processing techniques supported by background knowledge captured in domain-specific ontologies.

Keywords: Blog mining, blog search, information fusion, relevance, specificity, timeliness, credibility.

1. INTRODUCTION

Information retrieval (IR) has proven effective in the identification of authoritative sources on the web, but is the current IR model satisfactory for identifying authoritative blogs? In the following, we argue that it is not and propose alternative metrics for topical relevance ('aboutness'), specificity, timeliness, and credibility. Currently, we are prototyping these metrics in a blog search and analytics engine for news blogs on foreign relations and related topics.

Bloggers can be viewed as a loosely connected network of sensors closely monitoring news streams on specific subjects and posting commentaries or corrections, as they see fit. Just as with other networked sensors [11], in order to understand the situation that multiple sensors are (partially) reporting, care must be taken to aggregate what they are reporting correctly. Rigorous aggregation methods are especially crucial for events that trigger few sensors; rigor is less important when there is an abundance of evidence.

Currently, the most popular information retrieval (IR) models—beginning with Google's PageRank [6] and Kleinberg's HITS [14] algorithms—rank documents in terms of the frequency and position of query terms and the quality of the incoming (and outgoing, in Kleinberg) links. Unfortunately, IR models that work on ordinary web documents don't work well in the blogosphere because blog postings are short, rapidly updated, highly exophoric, highly quotational, and less susceptible to PageRank/Kleinberg analyses because they have few incoming links.

In our investigaton, we are concentrating on international 'news blogs', i.e., blog postings that are mainly commentaries on international news stories. Individual posts infrequently link to one another; they converge instead on the news stories they reference. Unlike the general Web where hyperlinks provide connectivity from any one document to another in 19 hops [2], weblog posts are largely unconnected. A blog may contain links to other blogs (a behavior called 'blogrolling'); but these links change infrequently and do not accurately reflect which links are currently considered most important [3]. A more direct referencing mechanism is the 'trackback' [21] where a blogger comments on another author's blog post and provides a link to it. Still, trackback links are less common in news blogs than links to target news stories. A recent study [1] showed that 18.2% of weblog posts in the news blog domain contain trackbacks, but 57.3% cite a news article, about 3.1 times more often. News stories themselves rarely contain hyperlinks to outside source material within the body of the story. The time frame for attention to a news story is also

rather short. As Dezso et al [9] show, the average "half-life" of a news story—the period over which it attains half of the clicks that it will ever acquire—is 36 hours.

2. INFORMATION RETRIEVAL METRICS

In applying IR analyses to mining newsblogs, it is crucial that we represent the ontology of news correctly since newsblogs are intimately connected with the news stream. *News events* are things that happen in the world and are reported in *news stories*. News stories are individuated by their author or source (e.g. a wire service), lede and publication date. The same story may appear in multiple *web documents,* or *URLs*, particularly stories of general interest and stories from wire services and major newspapers that syndicate their content. Like all events, *news events* are ontological individuals. News stories—and the blog posts that discuss them—are discrete textual units, so let us treat both news stories and blogs posts as individual documents.

We are interested in clustering or aggregating information about news stories and the blog posts that refer to them at the news event level (and, secondarily, at the story level). The URL level is not particularly interesting to us since we are interested in what a news story or a blog post is about and not primarily what website published the information, especially if other websites have precisely the same information.

In this section, we apply standard IR assumptions to blogs to see how these assumptions hold up. Consider the standard IR assumptions about web documents:

ABOUTNESS: In standard IR, what a document is about is usually modeled by the frequency of all the terms in the document (perhaps boosted by their position in titles and headings) divided by some factor corresponding to the (log of the) number of documents in which those terms appear: the so-called *tf*idf* (term frequency * inverse document frequency) *model* [12].

SPECIFICITY: Generally, the number of documents containing a term is inversely proportional to the specificity of a term. Therefore, specificity is modeled by the rarity of terms in a document. Corpus-wide, a document is most specifically about its rarest terms, and a document with only common terms is not specific.

CREDIBILITY: The credibility of a web document is a function of its incoming links. An important post will have many inlinks; an unimportant post will have few. If a post is high quality, the number and quality of documents linking to it will reveal its value over the long run.

ENDOPHORICITY: The context of a document is determined by the text of the post itself (i.e. it is *endophoric*). In classic IR, the text of a document is the best query for retrieving itself and thus a good way to retrieve documents similar to it. Outgoing links determine only the authoritativeness of the post, not what it is about. Incoming link text may be considered, but the text of linked documents is not considered.

While the above assumptions largely hold for documents like academic articles, blog posts tend to violate all of them. Consider a typical blog post of this type: "Abu Hamza wins right to appeal", (See Appendix A below) posted July 28, 2006 on the UK blog *This Scepter'd Isle* (sceptered-isle.blogspot.com) 1.

Based on the standard for term frequency, the blog post "Abu Hamza wins right to appeal" is not at all about *Court of Appeal*, *cleric*, *imam*, *Sir Igor Judge*, or *Abu Hamza al-Masri*, since none of these terms[1] appear in the blog post, only in the article to which it links. But intuitively that is not correct, and it would be wrong to judge this blog post as completely irrelevant to these topics. Nor would one want to say that the post is more about *(Edward) Fitzgerald* (2 mentions) than *London* (2 mentions), since presumably there are more posts about *London* than *Fitzgerald* (*pace* tf*idf). Finally, we would not want to say that this document has no credibility because it has no incoming links, and that the hyperlink to the BBC article serves only to indicate what the linked article is about via the text *appeal*. Given these intuitions about the standard IR model, let us reconsider the forms information retrieval metrics for blog post.

[1] Throughout, we will assume that we have a tokenized the texts into an optimal set of terms.

3. ABOUTNESS AND ENDOPHORICITY

First, we propose dropping the inverse document frequency factor as determining aboutness. What would the relevant document set be for which we are computing document frequency? Using the entire news or blog corpus would be misleading because it would be incorrect to penalize some topics for being frequent. News outlets in the US write about (search engine) Google about four and a half times as often as they write about (cell-phone maker) Nokia, but that doesn't mean that an article about Google and Nokia has to mention Google four times more often than Nokia for it to be equally about both of them. Similarly, if we *only* consider inverse document frequency (as in [13]), a wire story that is picked up by many outlets will be less about a person mentioned for every outlet that prints the story. That is clearly the wrong effect.

So, let us drop the inverse document frequency component as a factor of aboutness: a document is about something it mentions to some extent, no matter how many other documents talk about the same thing. As we have said, most of the blog posts we are interested in are commentaries *about* the news stories they link to. Our representation of their aboutness should reflect this.

It is therefore crucial to consider how news articles are structured. Most news stories today are written in the inverted pyramid style. To a first approximation, this means that the first (lede) paragraph says what happened (who, what, when, where, possibly why). The second paragraph then usually identifies the participants in the story more explicitly. Each subsequent paragraph contains some further fact that adds to the reader's comprehension of the central event, in order of decreasing importance. This is Journalism 101, but these facts of journalistic practice are not usually utilized in natural language processing or IR, except to the extent that terms are boosted for appearing at the beginning of articles. Clustering blog articles by the *news event* (not the URL) that they reference is crucial in newsblog analysis. A news story is about a particular news event (or, if it has multiple ledes, one event per lede), and events are individuated by their type and participants. Therefore, the proper names and event types in the lede paragraph (or that of a secondary, 'shirttail' lede) serve as an excellent representation of a news event (relative to a particular date).

Consider this typical example. In early November 2006, a certain Pakistani cleric was beheaded in the Waziristan region of Pakistan. Suppose you wanted to find all the blog posts about this news event. We have identified six blog posts in total in the blogosphere that call attention to this event. One of the posts doesn't identify the cleric by name. Some of the posts link to an Associated Press (AP) story (Appendix C) on the event and others to a Reuters story (Appendix D). Another links to an Al Jazeera story, and the fourth story linked to is in the *Qatar Peninsula Online*. The Reuters account and the AP account both render the name of the victim differently. Thus, searching on the name of the victim will not return all six of the blog posts. However, a search on the proper names and matrix verb in the lead of the Reuters story—i.e., Pakistani Taliban US Waziristan Afghanistan beheaded—and restricted to the date of event yields 73 URLs in the Google News index including the AP, Reuters, and Al Jazeera accounts that the blogs linked to (but not the *Qatar Peninsula* account, which uses the variant spelling 'Taleban' rather than 'Taliban'). This example illustrates the importance of clustering blog posts by the news event, not the URL that they link to. Aggregated by URL, each of these blog posts is about something different; aggregated by news event, we get a much more accurate sense of the reaction of the blogosphere to this event as a whole.

Even blog posts that link to a news article often do not quote enough of the article's important keywords to be findable on that basis. For example, in examining the blog posts that reference the top news story on blog monitoring service BlogPulse[2] over a 10-day period, a total of 211 blog posts are found. Of these, 49 (23.2%) don't contain any of the proper nouns or numeric entities in the lede paragraph of the news story they reference. Only 62 (29.4%) contain all of the proper nouns mentioned in the lede paragraphs. This sample population demonstrates the exophoricity of blog posts: they often don't explicitly mention the individuals involved in the event under discussion. What they are about must be discovered by way of the stories they reference.

[2] http://www.blogpulse.com

The inverted pyramid style suggests that the article (and, in a derivative way, the post that points to the article) is primarily about the objects, actions, and participants that are mentioned in the lede (and supporting graf) of the news story. Secondarily, it is about the elements mentioned in the rest of the story (or before the next subordinate or shirttail lede): In our *Hamza* example, the post is somewhat about the named entity *Edward Fitzgerald*, to a lesser extent *Sir Igor Judge*, and to a far lesser extent *al-Qaeda* and *Yemen* mentioned only in the last paragraph.

Formally, this suggests that a news story *d* is about a term *t*

$$\text{About}(t,d) = \sum(\alpha * \text{freq}(t)/\text{paragraph-rank})$$

Where paragraph-rank = 1 for the lede (first) paragraph, 2 for the supporting (second) paragraph and so on, and α is a scaling coefficient.[3]

Table 1 shows the difference in term prominence by aboutness and tf*idf for the BBC article 4.

Table 1. Tf*IDF vs. Aboutness

Term	Rank (tf*idf)	tf*idf[4]	Rank (About-ness)	Aboutness
Abu Hamza	1	18.46353	1	5.972222
al-Masri	7	2.968966	2	1
Court of Appeal	9	2.482951	3	1
London	6	3.009687	4	0.952381
Edward Fizgerald	3	8.422511	5	0.733333
US	12	0.289815	6	0.348485
New York	11	0.542623	7	0.285714
Igor Judge	2	9.0351	8	0.25
Finsbury Park	4	4.003229	9	0.222222
Old Bailey	5	3.22346	10	0.222222
Yemen	8	2.56096	11	0.2
al-Qaeda	10	2.088941	12	0.2

We submit that this representation of the news event is a better representation for finding other stories about the same event than a tf*idf representation. The tf*idf representation is optimal for finding other instances of the same story (at different URLs) while story's internal ordering of objects, actions, and participants distinguish the news event.

Following this line of thought, a blog post that hyperlinks to a news story is about what that news story is about, but to a somewhat lesser extent. The intuition here is that a blog post that was essentially just a hyperlink ("Read this!" or "All lies!" or "I can't believe this happened!") to a news story is *about* what it links to in some measure, but we must make certain that a blog post does not outrank a news article on a topic simply because it contains one more mention of a relevant term. Similarly, a blog post that links to another post that then links to an article is less about the article than the nearer post is. Let us then compute the aboutness of a term in a linked post inversely in proportion to the number of jumps from the original article. In this way, the aboutness due to outgoing links decays along the hyperlink graph.

---

[3] For now, we will assume all news articles are written in the inverted pyramid style. Of course, different models would apply to other styles.

[4] Document frequency figures were derived from Google News (news.google.com).

For an adequate measure of relevance, another feature of blog posts that must be accounted for is the extensive use of quotation. We have three intuitions. First, a post's *unselective* quotation of an entire news article does not make a blog post more about the subject of the news article than a post that merely hyperlinks to it. Unselective quotation does not add anything to the aboutness derived from hyperlink connectedness. On the other hand, a blog post that *selectively* quotes from an article is more about the elements within the selective quotation than a post that merely hyperlinks to a source article. That is, you can get more aboutness by quoting some than by quoting all, since selection adds information. Thus, raw count-based measures of aboutness in posts must take selective quotation into account. However, a post that selectively quotes just the prefix of an article (the first *n* lines) functions like the hyperlink as well; it contributes nothing more to the aboutness of the post that the original story does.

Let us regard selective quotations as amplifying what the original article said about a particular subject. Therefore, we will simply add the normalized frequency scores of quoted material to the *tf* scores from the original, applying the same decay function for additional paragraphs.

$$\text{About}(t,\text{quotation}) = \sum \left(a * \frac{\text{freq}_{\text{quotation}}(t)}{\text{paragraph-rank}_{\text{quotation}}} \right)$$

So, a post that selectively quotes some part of an original article can be more about the elements that feature in that quote than the original post itself, which seems intuitively correct.

Finally, blog posts make their own contribution to the discussion about a term; and, here, we would like this contribution to be proportional to the number of times a term is mentioned (*tf*). Blog posts provide commentary that does not follow the inverse pyramid structure; thus, it does not make sense to decay for position. Also, we do not use inverse document frequency to measure aboutness, as discussed above. Blog post aboutness is then just the sum of the aboutness of the hyperlinked document, the aboutness of any quotation, and the aboutness of original content.

4. SPECIFICITY

Inverse document frequency was originally seen as a way of increasing the chance of returning a relevant document by emphasizing rare query terms over common ones. It was seen as a proxy for term specificity [19]. However, the distribution of news stories and the unequal coverage of news events call document frequency into question as a measure of news specificity. Specificity in a blog post, like aboutness, is derived from both what the post references and from the original content of the blog post itself. Specificity has several aspects: descriptive specificity, grammatical specificity, and ontological specificity. The challenge is to make these metrics commensurable so that they can be combined linearly.

Specificity applies to stories (or documents). News stories are by their nature specific; to a first approximation, they report on an individual event (one per lede) and then provide some context to it. But how is the uniqueness of this event quantified in the stories themselves? It can't be modeled in terms of the number of stories (or URLs) in which a term appears. There are stories every day about, say, the *US* and *Iraq*. But, unlike the old Chevy Chase faux news bulletin [20], they do not thereby fail to be news. It is their novel combination of elements that is newsworthy. So while two stories are about the same news event if they refer to an event with the same participants at the same time and place (see [15]), to a first approximation, one story is more specific than another (on the same news event) if it contains more (specific) proper names, dates and amounts.

For example, referring to our beheading news story, the following sentence:

(1) A Muslim cleric, Maulana Silahuddin, was found beheaded in the Razmak area of North Waziristan.

is more specific than

(2) A Muslim cleric was found beheaded in Waziristan.

Both sentences are about the same event, but the first sentence is more specific because it contains more named entities. This suggests, then, that the specificity of a news story is primarily a function of its length, which is just what the inverted pyramid model suggests: paragraphs further down the story provide more and more detail or background.

News stories, after all, are not blind items: the participants and events must be named (or at least characterized as much as possible, in light of protecting sources).

So, a simple measure of specificity of a new story *d* would be:

$$\text{Specificity}(d) = |\{\text{proper names, measures}_d\}|$$

This crude metric is quite effective (although it helps to count reduced forms of names as instances of their full form). For example, in the AP and Reuters accounts of the beheading in Appendix C and D, the AP account is somewhat more specific, having 21 unique named entities while the Reuters account has only 17.

To refine this metric further, consider cases like this:

(3) A Muslim cleric was found beheaded in Pakistan.

Here, sentence (2) and (3) are tied for number of unique proper names, but (2) is more specific than (3). While the unique proper name count metric is straightforward and somewhat useful, unique proper names should be weighted by their specificity if an ontology[5] of the subject matter is available (see [16] for work along similar lines). Thus, sentence (2) would outweigh (3), as Waziristan is a region within the country of Pakistan and, therefore, the more specific term.

$$\text{Specificity(document)} = \sum \text{(named entity * specificity(named entity))}$$

Finally, grammatical specificity should be incorporated: a phrase such as *a White House source* is less specific than a proper name (e.g. *Scooter Libby*), so a special discount for proper names embedded in indefinite noun phrases (including plurals) is justified.

Extending this metric to blog posts is a simple matter. A blog post's specificity is a function of the number of unique named entities (and their ontological and grammatical specificity) in the combined text of the blog post text and the news story it references.

5. TIMELINESS

There is both a retrospective (e.g. "a timely response") and a predictive (e.g. "this is a timely article") sense of timeliness. Both are instances of a single, unified sense: a text is *timely* when it is published close in time to the salient event it discusses. In the retrospective sense, a response is timely if it comes soon after the event it answers; in the predictive sense, a remark/article/conversation is timely if it is produced (or received) close in time to a relevant event it precedes (e.g. predictions of market instability were timely in pre-crash 1929; predictions of market stability were no.).

Let us focus primarily on the retrospective sense since newsblogs are mostly about events in the (recent) past. Timeliness in the retrospective sense should not be mistaken for recency, a post's proximity to the present. Timeliness is more important. As noted in (9), more than half of the visits that a news article gets will be made in the first 36 hours after it is released. Similarly, thirty-six hours is the half-life of news articles in the blogosphere. Therefore, in ranking blog entries on the topic of a news story, it is more representative to know what people thought in the immediate wake of the event than much later. Later posts are often restatements or endorsements of others' views.

Timeliness can be computed as the difference between the time of publication of a blog post and the time of publication of the source article it references. If a news source doesn't have an RSS or Atom feed, it will be necessary to parse a news source text to identify the most likely date of publication. In the case of our BBC article, the article is

[5] Assigning a specificity to a term in an ontology based on its depth or height is not straightforward in practice. Deployed term hierarchies are seldom designed so that all concepts at one hierarchical level are ontologically at the same level. This is not just because of sloppiness. First, Different branches of the hierarchy can be elaborated with different amounts of detail depending on the importance of the branch. Second, one can define new classes for a variety of purposes. Such classes may or may not have proper names. Third, one can infer classes, so new classes can appear in the hierarchy even without an explicit declaration. All of these reasons make "depth" a poorly defined notion in an ontology, so it is problematic to employ such a notion here. The same would be true in terms of 'height' as measured by the number of subclasses or part-whole relations in the hierarchy.

timestamped with the line *Published: 2006/07/28 15:36:45 GMT*. The blog post has an associated publication timestamp of *Fri, Jul 28 2006 1:31 PM***.** In order to compare dates, we will need to normalize dates into a common format over which date arithmetic can be performed. Notice that in this case, the blog post seems to precede the publication of the article it cites, so doing the necessary date arithmetic would require more than just the timestamp, but would crucially depend on identifying the time zone of the blogging site as well, if not the author.

Locality is the spatial dual of timeliness: how physically close is the blogger to the events he or she links to. Since many blogs are hosted on blogging services, there is no standard part of a blog feed that encodes where a blogger is geo-located. This information is frequently given in blog profiles, however. In our news article, there is no physical dateline mentioned, either, although the URL for the news story does contain the string 'uk_news'. So, again, computing the locality of a blog post to its associated event requires some sophistication and reasoning.

6. CREDIBILITY

Credibility is difficult to measure objectively. Someone is credible if we have evidence that what he or she says is accurate or persuasive. Credibility judgments can change over time. Independent assessments of the accuracy of a source can inform one's credibility judgments, but lacking credibility doesn't mean that a source can't speak the truth. It just means that one's perceived veracity is low.

In the standard IR model, credibility is a function of inbound (PageRank) and outbound (HITS) links. A page is of good quality, and therefore presumably credible, if good pages point to it, and/or if it points to good pages. This standard is used by at least one prominent blog search engine (Technorati) to calculate blog authority directly as (the log of) the number of incoming blog links over a six month period [18]. A PageRank-style algorithm would go beyond this and calculate a blog's goodness recursively, weighing the number of links from highly inlinked blogs more heavily, and so on. However, blog credibility metrics that consider only aspects of the reception of a message are obviously one-sided. Reception-only metrics consider a blog with no inlinks to have no credibility. Blogs, like other messages, however, have a source, content and receiver(s), and all three can contribute to a blog's (perceived) credibility [7].

The problem with considering only the reception of a blog in credibility attributions is that it tends to unfairly reward blog longevity. As David Sifry's slide on blog authority illustrates [17], high inlinking correlates with blog age and post volume. Measured in this way, new bloggers will always tend to have less credibility. Sifry is right to say that inlink counts are perhaps a better measure of blog *influence*.

Are credible bloggers always influential? As an experiment, we examined the list of bloggers labeled *Political Science Weblogs* on the popular academic link aggregator *Political Theory Daily Review* (politicaltheory.info) run by doctoral student Alfredo Perez. Perez's list of academic political scientist bloggers from across the ideological spectrum hasn't been updated in a while, but 25 of the 27 bloggers he lists have active blogs. These bloggers has collectively published at least 62 books and approximately 650 peer-reviewed articles. There are 18 tenured faculty members and 21 PhDs. But even including the very popular blogs danieldrezner.com and angryarab.blogspot.com, the median blogger in this list has links from only 23 blogs in the last 6 months, so only counts as having middle authority. Thus, having demonstrable credibility by measures of a profession is no guarantee of inlinking popularity. Further, it is unclear that one tenured political science professor is up to four times as credible as another just because one's blog is many times more popular than the other. The less-popular professors might rightly claim to be aiming for more intellectual rigor, but being less accessible shouldn't make them less credible.

Another example vividly illustrates why blog popularity should not be equated with credibility. Recently, we ran a query on Technorati[6] on the Gospel of Judas, the newly discovered Gnostic gospel published in 2006, and sorted by authority. The top result for the query [gospel of judas] was a post by one "Zeus Bhagwan" on the "citizen journalism" site *News is Now Public*[7]. The entire site had 877 inlinks, but there were no trackbacks for this particular post. In his profile on nowpublic.com, Zeus Bhagwan describes himself as follows [5]:

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

[7] http://www.nowpublic.com

This post was his only pronouncement on the subject, and in fact, his only post ever. Contrast Zeus Bhagwan's record with Prof. James R. Davila of the University of St. Andrews, Scotland, whose blog PaleoJudaica [8] had only 128 inlinks at the time. Prof. Davila has a PhD in Near Eastern Studies from Harvard and has published at least 3 books and over 20 peer-reviewed articles on early Christian writings and early rabbinic Judaism. Prof. Davila has published 61 posts on the topic of the Gospel of Judas alone. By ordinary measures, Davila would seem to be the more credible commentator; but, considering only blog inlinks, Zeus Bhagwan is 140% more authoritative than Davila.

We have constructed a measure of blog credibility that takes into account *source*, *message* and *reception* features of bloggers. As the Zeus Bhagwan case illustrates, it is important to construct credibility metrics per author, rather than per blog, because in a group blog situation, an author can derive authority he or she hasn't earned by posting to a popular blog.

Source features relate to the signals of identity, expertise, and reliability [10] that an author places to underpin his or her credibility. These signals include blogging under a real name or providing a real name in a blog profile, posting an email address, posting a PayPal or similar account, posting organizational affiliations, posting location, linking to a resume or CV, posting photos of oneself, and so on. All of these partial disclosures of identity serve to make an author more credible by making the author accountable for what he or she writes or claims to know. Many bloggers have a *nom de blog*, but they often reveal quite a bit more about themselves in their blog profiles. On most blogging platforms, there is a convention for where the blogger profile is found in relation to the main blog. On Typepad blogs (www.typepad.com), for example, the blogger profile page is located at about.html under the blog's main URL.

Blogger profiles contain a wealth of information that can be important in evaluating a blog. This information is largely ignored by blog analytics systems today. A large number of blogs provide a geographic location for the blogger that could be used to determine whether a user's posts represent on-the-ground reporting or second-hand knowledge.

Other source-related features are analogous to signal strength: the number of posts an author writes per month, the number of topics addressed (as measured by post tags or unique proper name counts), and the ratio of original text to ads (a feature of spam blogs).

Significant message features include the number of original (non-quoted) sentences in a post, the presence of a non-default tag or category assigned to the post, the mention of a source publication, the inclusion of a source URL, the (log) circulation/readership/viewership figures for the source publication (available from the Audit Bureau of Circulation online), whether the source publication has an associated ISSN or ISBN number, the inclusion of a documentary photograph or video link (such as a YouTube video), writing from personal experience, mentioning proper names not in a quoted source, lack of spelling/grammar errors and profanity, and so on.[8]

These message features can be averaged over the entire blog, recent posts, or just those posts that contain certain keywords in order to provide overall, recent or topic-specific measures of message content.

Finally, reception specific features include the (log) number of trackbacks (links from other blogs), the number of comments using CAPTCHA (human-detection technology), the number of non-CAPTCHA comments that don't include a hyperlink (a comment spam necessity), the number of subscribers to the blog, the number of hyperlinks in social bookmarking services (such as del.icio.us, reddit or digg[9]), and so on.

Currently, we have a set of approximately 50 such source, message and reception features that we are using to measure blog author credibility.

---

[8] Sophistication is called for here. If a post cites an external news source with poor credibility in order to refute or criticize it, the poor reputation of the source shouldn't taint the credibility of the blog.

[9] http://del.icio.us, http://www.reddit.com, http://www.digg.com

Using these metrics, it is clear that an author like Zeus Bhagwan would be assigned almost zero credibility, since he has just one post (albeit on a popular blog), he provides no verifiable information about himself or contact information indicating accountability, he links to no known source material, and his post has made no impression in the blogosphere: there have been no comments, trackbacks or bookmarks. This doesn't meant that Zeus Bhagwan's post is not true; but his post does lack external factors providing reasons to believe it or take it seriously. His post contrasts with Prof. Davila's blog, which includes a large number of the features mentioned. Davila's posts on this and other topic thus would have a large degree of credibility on our metric.

As we go further, we will determine which of these features are the most significant or informative in correlation with perceived credibility. We will then determine which of these features can be identified (or approximated) automatically, and thus determine what automatic measures of blog credibility are possible and how close they come to ideal measures of blogger credibility (as objectively assessed by humans). In a small experiment along these lines, three colleagues provided a partial ranking of 10 blogs in terms of credibility selected by querying on the topic [Taliban]. Human rankings had an average pairwise Kendall tau rank correlation of 0.51 (where 1.0 indicates perfect correlation and –1.0 indicates perfect uncorrelation). The average pairwise Kendall tau correlation of humans with the assigned credibility metric ranking was 0.45. When compared with the rankings determined by inlink counts (as determined by Technorati), the average pairwise Kenall tau correlation with human rankings was only 0.30. This is an encouraging result; it suggests that human credibility judgments are correlated with features in addition to inlink counts. Of course, further work is necessary to determine which features are worth tracking. In this experiment, the credibility metric used was just the sum of all the factors mentioned above, weighted equally. In future experiments, the factors could be weighted differentially in order to achieve greater correlation with human judgment. However, since the inter-annotator agreement here was not high (0.5), it would be better to train the weights against a larger dataset with greater inter-annotator agreement.

It is important to reiterate that we are only considering objectively determinable features in this measure; we are not including any subjective features. It is our hope that a subset of these features is computationally tractable and can be deployed to measure blog credibility in a sophisticated, recursive PageRank-like algorithm that assigns measures of blog credibility on the basis of inherent blog features and the credibility of blogs that have inbound links to the blog under consideration.

7. CONCLUSION

By examining the nature of newsblogs and the news stream they monitor, we have proposed more suitable metrics for evaluating the topical relevance, timeliness, specificity and credibility of newsblog postings than the current standard model. In all cases, it is important to consider the news article to which a blog post links as well as the blog post itself in computing these metrics. Implementing these models effectively involves shallow, automatable analysis of blog post content. In this analysis, it is important to distinguish original post content from quotations, and links to news stories from links to other blog posts. The effort involved is warranted by the increasing importance of blogs in both shaping and reflecting discourse about important events.

Appendix A: Example Blog Post 1
ABU HAMZA WINS RIGHT TO APPEAL

Abu 'The Hook' Hamza has been granted the right to ***appeal*** sentence for inciting racial hatred and soliciting murder. More public money will now be poured down the drain to help Hamza win his freedom.

<quote>Mr. Fitzgerald said these events included the 11th September attacks on New York, and the 7th July bombings in London. "It further meant that he was subjected to a relentless campaign of adverse media publicity condemning him as a preacher of hate and inciter of violence," he said**.** </quote>

What about the fact he actually delivered hateful sermons and clearly incited violence? He praised the 'Magnificent 19' suicide terrorists who killed over 3000 people in New York and Washington. He called for jihad and the London bombings took place. How outrageous that his defence lawyer tries to suggest that it was only these events and the wider coverage of Hamza's vicious and hate filled preaching that have made him look bad.

The only crumb of comfort is that if the British judiciary set this terrorist recruiting sergeant free the American and Yemeni authorities will want a piece of him. Why do we tolerate people like Hamza in this country? The 'Londonistan' tag coined by the French is so appropriate.

Appendix B: Article cited by Post 4
JUDGES APPROVE ABU HAMZA APPEAL

Islamic cleric Abu Hamza al-Masri has been given the go-ahead to challenge his convictions.

Lawyers for Abu Hamza, 48, told the Court of Appeal that the long delay in bringing a prosecution against him had made a fair trial "impossible".

Judges at the London court said the case was "arguable" and agreed to a full hearing in October. Abu Hamza, who did not appear in court, was jailed for seven years for inciting racial hatred and soliciting murder.

'Adverse publicity'

Edward Fitzgerald, QC for the preacher, told the court that his client was convicted on the basis of speeches he had made between 1997 and 2000 - six years before his trial.

"This meant that a unique series of events supervened which prejudiced his chances of a fair trial."
Mr. Fitzgerald said these events included the 11 September attacks on New York, and the 7 July bombings in London.

"It further meant that he was subjected to a relentless campaign of adverse media publicity condemning him as a preacher of hate and inciter of violence," he said.

Martyr claims

Sir Igor Judge, sitting with two other senior judges, said he did not wish to "raise any false optimism" for Abu Hamza.

He added: "But in our judgment there are a number of grounds drawn to our attention which are arguable."

Abu Hamza, a former imam at a mosque in Finsbury Park in north London, was found guilty by an Old Bailey jury of 11 charges - including six of soliciting murder.

Following the conviction, his lawyer said he considered himself "a prisoner of faith" subject to "slow martyrdom".

He is also wanted in the US on charges of providing support to al-Qaeda, and over his alleged involvement in a hostage-taking conspiracy in Yemen.

The US is seeking his extradition but has agreed to postpone proceedings until after his appeal has been resolved.

APPENDIX C: AP ACCOUNT[10]
TEACHER BEHEADED FOR SPYING ALLEGATIONS
(21 Unique named entities underlined)

By Bashirullah Khan

Miran Shah - Pro-Taliban militants beheaded a Muslim seminary teacher after kidnapping him over suspicion that he was spying for the US forces in neighbouring Afghanistan, residents and an official said on Friday.

Maulvi Silahuddin, a teacher at a madrassa or Islamic religious school in the North Waziristan tribal area's Nara Bakakhel village, went missing on Thursday, and villagers spotted his body early on Friday near a ditch in the area, said a resident, Gul Janan.

"I saw his headless body. He had also been shot in the chest," Janan told The Associated Press by telephone.

[10] http://www.iol.co.za/index.php?set_id=1&click_id=123&art_id=qw1162539721846B212

A letter found nearby read: "Anyone spying for America will face the same fate," a local security official said on condition of anonymity because he was unauthorised to speak to the media.

Nara Bakakhel lies 75 kilometres (45 miles) south of Miran Shah, the main town in North Waziristan, a volatile tribal region bordering Afghanistan. The government in September signed a peace deal with pro-Taliban militants and elders to end violence there.

Although the truce is still holding, local militants still sometimes kill people suspected of spying for the Pakistani or US forces.

Pakistan is a key ally of the United States in its war on terror, and it has deployed about 80 000 troops in its semi-autonomous tribal regions, from where Islamic militants are believed to cross over to Afghanistan to target Western forces. - Sapa-AP

APPENDIX D: REUTERS ACCOUNT[11]
PAKISTANI TALIBAN BEHEAD "US SPY"

(17 unique named entities underlined.)

Pakistani Taliban fighters beheaded a tribal cleric accused of being a US spy in the Waziristan region bordering Afghanistan, a security official in the restive tribal region said.

The body of Maulana Salahuddin, 45, was found on Friday on a road between North [Waziristan] and South Waziristan, two semi-autonomous regions regarded as hotbeds of support for al Qaeda and the Taliban fighting NATO, US and Afghan forces across the border.

A note pinned to the cleric's body described him as an American spy, the security official said. The corpse was sprayed with bullets after the beheading, he added.

Pakistan's government signed a pact with tribal leaders in North Waziristan on September 5 to end clashes between pro-Taliban militants and Pakistani security forces.

Since the deal was reached, US military officials say, attacks against US-led NATO troops and Afghan government forces have tripled.

Pro-Taliban tribesmen appear to be violating the pact also by setting up a parallel administration in North Waziristan, just as they did after a similar treaty in South Waziristan.

ACKNOWLEDGMENTS

This material is based upon work supported by the United States Air Force under Contract No. FA9550-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

REFERENCES

1. 'Monty'. 2006. This Scepter'd Isle blog: Abu Hamza Wins Right to Appeal. http://sceptered-isle.blogspot.com/2006/07/abu-hamza-wins-right-to-appeal.html.
2. Albert, R.; H. Jeong, and A.-L. Barabási. 1999**.** Diameter of the World Wide Web. *Nature* 401, 130-131 (1999).
3. 'Atrios'. 2007. Eschaton blog: Blogroll Amnesty Day **http://atrios.blogspot.com/2007_01_14_atrios_archive.html#116879512760619699**
4. BBC. Judges approve Abu Hamza appeal. http://news.bbc.co.uk/1/hi/uk/5224978.stm.
5. Bhagwan, Zeus. Profile. http://members.nowpublic.com/zeusbhagwan
6. Brin, S. L Page, R. Motwami and T. Winograd. 1999. The PageRank citation ranking: Bringing order to the Web. Stanford University Technical Report 1999-0120.

[11] http://tvnz.co.nz/view/page/411319/879286

7. Danielson, D.R., 2006. Web Credibility. In Claude Ghaoai, (ed.) *Encyclopedia of Human-Computer Interaction*. Idea Group. Hershey, PA. 2006.
8. Davila, JR. PaleoJudaica blog. http://Paleojudaica.blogspot.com
9. Dezso, Z., E. Almaas, A. Lukacs, B. Racz, I. Szakadat, A.-L. Barabási *Dynamics of information access on the web* Physical Review E 73 (6): Art. No. 066132 Part 2, (2006)The IDF Page. www.soi.city.ac.uk/~ser/idf.html
10. Donath, J. 2000. Being Real. In K. Goldberg (ed.) *The Robot in the Garden: Telerobotics and Telepistemology in the Age of the Internet* Cambridge, MA: MIT Press
11. Hall D.; J. Llinas, 1997. An introduction to multisensor data fusion, Proceedings of the IEEE Vol. 85, No. 1, pp. 6--23, January 1997.
12. The IDF Page. www.soi.city.ac.uk/~ser/idf.html
13. Ikeda, D., T. Fujiki, M. Okumura, 2006. Automatically Linking News Articles to Blog Entries, AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Palo Alto.
14. Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46, 1999.
15. Parsons, T. 1994. Events in the Semantics of English: A Study in Subatomic Semantics. Cambridge: MIT Press.
16. Shin, HW, Eduard Hovy, Dennis McLeod, Larry Pryor**.** 2005. *Generality: A New Criterion for measuring generality of documents.* USC CS Technical Report 05-849.
17. Sifry, D. 2006a. Blogging Characteristics by Technorati Authority. http://www.sifry.com/alerts/Slide0006-8.gif
18. Sifry, S. 2006b. State of the Blogosphere. October, 2006. http://www.technorati.com/weblog/2006/11/161.html
19. Spärck Jones,K 1972 A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**, 11-21, 1972 and **60**, 493-502, 2004
20. Wikipedia. 2007. Generalissimo Francisco Franco is Still Dead. http://en.wikipedia.org/wiki/Generalissimo_Francisco_Franco_is_still_dead
21. Wikipedia. 2007. Trackback. http://en.wikipedia.org/wiki/Trackback