

New Metrics for Newsblog Credibility

Brian Ulicny

bulicny@vistology.com

Ken Baclawski

ken@baclawski.com

VISTology, Inc., 5 Mountainview Drive, Framingham, MA 01701 USA

ABSTRACT

The blogosphere is an invaluable source of insight into attitudes towards significant world and local events. Traditional measures of topical relevance, timeliness, specificity and credibility are inadequate when it comes to blogs, however, due to their short length, high degree of quotation, exophoricity, and the short life cycle of blog postings. In this paper, we motivate a novel metric for blog credibility that is one of the metrics underlying a blog search and analytics engine we are constructing.

Keywords

Blog mining, blog search, credibility, authority, influence.

1. INTRODUCTION

Is the standard model of information quality satisfactory as applied to blogs about the news stream? In the following, we argue that it is not and motivate an alternative metric for blog post credibility that we are currently prototyping in a blog search and analytics engine for news blogs on foreign relations (see [11] for details).

The sort of blog postings we are interested in -- call them *newsblogs* -- are mainly commentaries on news stories. They largely consist of blog posts featuring hyperlinks to, references to or explicit quotations from traditional news sources, such as newspapers and news broadcasts. Many news events light up a large portion of the blogosphere (e.g. [1]) without being really consequential. How can we incorporate the credibility of commentators into an assessment of the importance of a news event in addition to popularity?

Credibility is difficult to measure objectively. Someone is credible if we judge them to be disposed to tell the truth [10] or present sound arguments on a topic. Credibility is distinct from reliability: we make judgments of credibility even when we don't have independent assessments of a source's reliability. Independent assessments of the accuracy or reliability of a source can inform one's credibility judgments. On the other hand, lacking credibility doesn't mean that a source doesn't speak the truth: it just means that the source is not likely to be believed or be found persuasive.

Currently, the most popular information retrieval (IR) models, beginning with Google's PageRank [2] and Kleinberg's HITS [7] algorithms, rank document relevance to a query *quantitatively* in terms of query term frequency and position

and *qualitatively* via a recursive metric of the quality of incoming (and outgoing, in Kleinberg) links. It would be natural to take high-quality blog posts to be credible. However, retrieval, clustering and indexing techniques that work on ordinary web documents don't work well in the blogosphere, because blog postings are short, of ephemeral importance, highly exophoric, highly quotational, and much less susceptible to PageRank/Kleinberg type analysis because they have relatively few incoming links, especially on a per-author basis.

2. CREDIBILITY vs. AUTHORITY

The Technorati¹ blog search engine calculates a measure of blog *authority* as (the log of) the number of incoming blog links over a six month period [9]. Authority would seem to be closely related to the notion of credibility. Is inlink quantity therefore a good correlate for credibility?

Blog credibility metrics that consider only aspects of the reception of a message (here, inlink counts) 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 credibility [4]. One problem with considering only the reception of a blog in credibility attributions is that it tends to unfairly reward blog longevity. As Sifry notes [8], high inlinking correlates with blog age and post volume. Therefore, new bloggers will always tend to have less credibility, measured in this way. Sifry is right to say that inlink counts are perhaps a better measure of blog *influence*: a notion that is essentially backward-looking.

Are credible bloggers always influential? As an experiment, we looked at the list of bloggers labeled *Political Science Weblogs* on the popular academic political theory link aggregator *Political Theory Daily Review*². This list of political scientist bloggers ranges across the ideological spectrum. It hasn't been updated in a year, but 25 of the 27 bloggers he lists have active blogs. These bloggers have collectively published at least 62 books and approximately 650 peer-reviewed articles. There are 18 tenured faculty members and 21 PhDs.

¹ <http://www.technorati.com>

² <http://www.politicaltheory.info/weblogs.htm>

The median blogger in this list has links from only 23 blogs in the last 6 months, so it only counts as having middle authority. Thus, having demonstrable credibility by external measures is no guarantee of inlinking popularity or authority. Further, it is unclear that one tenured political science professor is up to four times more credible than another just because one professor's blog attracts a few hundred more links than the other's.

3. MEASURING CREDIBILITY

We are constructing a measure of blog credibility that takes into account *source*, *message* and *reception* features of bloggers. We also consider it crucial to measure credibility by author, rather than by blog, because in a multi-author blog, an author can derive credibility he or she hasn't earned (e.g. see [11]). Currently, we have identified a set of 48 source, message and reception features that we are using to formulate a measure of blog author credibility. As with Google's news ranking algorithm [3], we rely on source and message features when inlinks are rare. Kaid and Postelnicu [6] have also demonstrated that both message features and source features factor into credibility assessments.

In evaluating a preliminary version of this metric, three colleagues provided a partial ranking of ten blogs in terms of credibility selected by querying on the topic [Taliban]. A credibility measure was computed using the unweighted presence or absence of the 48 factors mentioned. 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 Technorati inlink counts, the average pairwise Kendall tau correlation with human rankings was only 0.30. This was an encouraging result; it suggests that human credibility judgments are correlated with features in addition to inlink counts. Further work is necessary to determine which features are most highly correlated with credibility. By finding a set of bloggers who we can consider credible *a priori*, such as the tenured bloggers mentioned above, we can identify features that distinguish their blogs from ordinary blogs.

Along these lines, in a study of twenty blog posts from PoliticalTheory.info bloggers with academic tenure and a set of twenty random blog posts retrieved on [Taliban], it is clear that there are some features that distinguish credible bloggers from average bloggers (Table 1: see first five rows, italicized). Most textual features we examined (number of sentences, number of proper names, reading level) were fairly similar on average between the tenured bloggers and average bloggers, so writing style alone doesn't distinguish credible bloggers. However, a number of features did stand out as being significant, such as use of full names and affiliation, presence of comments, and hyperlink citations. We hope to use these factors to tune our credibility metric further

4. ACKNOWLEDGMENTS

This material is based upon work supported by the United States Air Force under Contract No. FA9550-06-C-0023. This material expresses the views and conclusions of the authors alone and

does not necessarily reflect the views of the United States Air Force.

Table 1: Some Distinguishing Features of Tenured Bloggers

	Tenured Bloggers	Average Bloggers
<i>Full Names</i>	100%	30%
<i>Affiliation</i>	100%	10%
<i>Unquoted Content</i>	100%	75%
<i>Links to News Sources</i>	90%	45%
<i>Has 1+ Comments</i>	60%	30%
Avg # Sentences	21.8	29
Avg # Paragraphs	9.9	10.4
Avg Unique PNs	18.3	16.7
Avg. Readability ³	15.1	13.6

5. REFERENCES

- [1] BoingBoing. 2007. "Jilted, diapered astronaut planned to kidnap rival".
- [2] Brin, S. L Page, R. Motwami, T. Winograd. 1999. The PageRank citation ranking: Bringing order to the Web. Stanford University Technical Report 1999-0120.
- [3] Curtiss, M, K. Bharat, M. Schmitt. 2005. Systems and methods for improving the ranking of news articles. US Patent Application 20050060312
- [4] Danielson, D.R., "Web Credibility", in *Encyclopedia of Human-Computer Interaction*. Claude Ghaoai, ed. Idea Group. Hershey, PA. 2006.
- [5] 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.
- [6] Kaid, L.L, M. Postelnicu. 2007. Credibility of Political Messages on the Internet: A Comparison of Blog Sources. In M. Tremayne (ed.), *Blogging, Citizenship, and the Future of Media*. Routledge: New York.
- [7] Kleinberg, J. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.
- [8] Sifry, D. 2006. Blogging Characteristics by Technorati Authority. <http://www.sifry.com/alerts/Slide0006-8.gif>
- [9] Sifry, S. State of the Blogosphere. 2006. <http://www.technorati.com/weblog/2006/11/161.html>
- [10] Thagard, P. 2005. "Testimony, Credibility and Explanatory Coherence". *Erkenntnis* (2005) 63:295-316
- [11] Ulicny, B., K. Baclawski, A. Magnus. 2007. New Metrics for Blog Mining. *Proceedings of SPIE Defense & Security Symposium '07* Vol. #6570. Orlando, FL.

³ Gunning fog index