Metrics for Monitoring a Social-Political Blogosphere

A Malaysian Case Study

The authors' automated framework evaluates blog posts in a social-political blogosphere and, by aggregation, entire blogs according to their relevance, specificity, timeliness, and credibility. These metrics are superior to current methods in information retrieval for blogs because they better reflect the distinctive hyperlink structure of a social-political blogosphere than do other methods. The authors chose the Malaysian social-political blogosphere as a case study because of the role Malaysian bloggers played leading up to that country's 2008 general election, and afterward.

Brian Ulicny and Christopher J. Matheus
VIStology

Mieczyslaw M. Kokar
Northeastern University

The value of any piece of information depends on its relevance, specificity, timeliness, and credibility to the task at hand. In recent years, local social-political (sopo) blogospheres have become an important arena for political mobilization, but qualitatively for an analyst or interested observer, it's difficult to evaluate information in an unfamiliar blogosphere. Here, we present novel metrics for monitoring an unfamiliar blogosphere and demonstrate their superiority over current metrics. We've implemented an automated framework for evaluating information in sopo blogospheres in a way that better reflects their structure. We chose the Malaysian sopo blogosphere as a case study because of the role Malaysian bloggers played leading up to that country's 2008 general election, and afterward.

We developed four independent metrics. To determine the *credibility* of a blog author's posts, we use authority (centrality), engagement (user comments), and various credibility-enhancing features that we've validated as informing human credibility judgments, such as blogging under a real name, listing affiliations, and blogging over a long time period. We determine a blog post's *relevance*, or what it's about, not only by its text but also by the text of any news article it references. *Timeliness*, as distinguished from recency, is about proximity to a relevant event. We determine a blog post's timeliness by comparing its time stamp with the

 1089-7801/10/$26.00 © 2010 IEEE

publication date of a news article that it cites. Finally, the number of unique proper nouns mentioned in both a blog post and any news article it cites determines its *specificity*.

The Malaysian Sopo Blogosphere

The Malaysian blogosphere provides an interesting set of properties for analysis. Although the Malaysian press is tightly controlled, the government has officially encouraged Internet-based enterprises and rejected Web censorship or filtering. (Article 3 of the 1998 Malaysian Communications and Multimedia Act states that "Nothing in this Act shall be construed as permitting the censorship of the Internet."[1]) Consequently, the Malaysian blogosphere has become an important locus for political opposition activity.

Since 2007, however, the Malaysian government has attempted to intimidate opposition bloggers via lawsuits, police interrogations, and arrests.[2] In August 2008, the Malaysian government blocked access to the online opposition news source *Malaysia Today*, and its publisher was jailed under the Internal Security Act.[3]

Nevertheless, on 10 November 2007, tens of thousands of Malaysians potentially risked prison sentences to participate in a rally for electoral reform in Kuala Lumpur (known as the Bersih rally), the first mass rally there in a decade. In the 2008 general election, the ruling party and coalition ultimately lost the supermajority they'd enjoyed since 1969 that had enabled them to modify the country's constitution at will. The Bersih rally couldn't have been mobilized without the Internet, aided, perhaps, by technologies such as SMS. Several bloggers went on to win seats in parliament.

Blogger.com, the most popular blogging platform in Malaysia, has more than 152,000 Malaysian profiles — many more than on Wordpress.com or similar services. The number of people using social networking sites, however, dwarfs the number of bloggers. As of August 2008, 5.4 million Malaysians had profiles on Friendster, 500,000 were on Facebook, and 293,000 were on MySpace.

To determine the size of the Malaysian sopo blogosphere, we performed a Web crawl seeded with the 385 blogs in the Sopo-Sentral directory — a self-reported directory of Malaysian sopo blogs (see http://sopo-sentral.blogspot.com) — at a depth of four (that is, we processed each URL's text, followed by the text of each page hyperlinked from that text, and so on for two further iterations). This crawl produced 220,320 unique URLs representing 4,693 unique sites. Of these, approximately 2,000 were blogs, including 1,060 on Blogger, 801 on Wordpress, 16 on Typepad, and the remainder either self-hosted or on other blogging services, such as Xanga, Vox, or LiveJournal. A later crawl at the same depth of 1,271 blog posts mentioning "Bersih" in the week following the rally revealed 878 unique blogs, with 408 on Blogger, 228 on Wordpress, and the rest on other services.

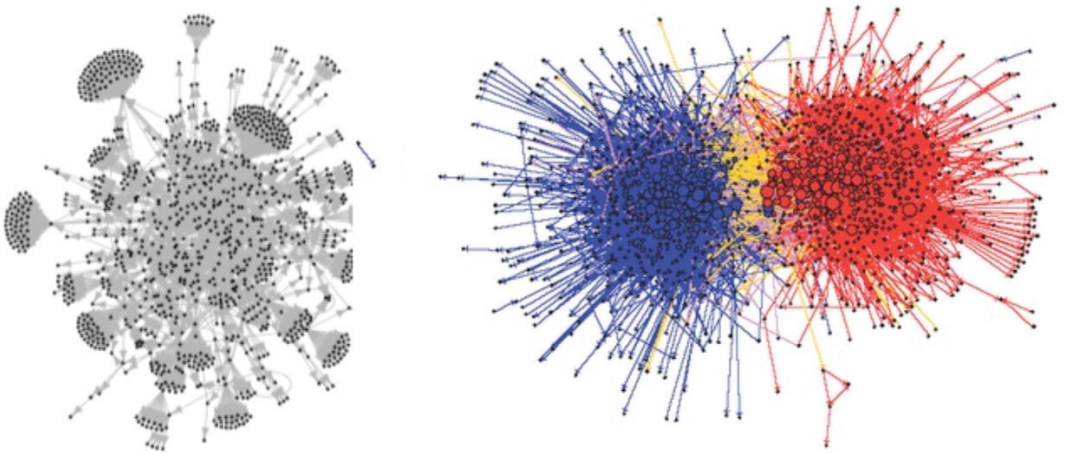
Figure 1 shows the Malaysian sopo blogosphere's link structure based on our original crawl. Unlike Lada Adamic and Natalie Glance's depiction of the US political blogosphere, little evidence exists of polarization in the Malaysian structure.[4]

Related Work in Monitoring International Blogs

Although blogging has drawn considerable attention from both computer and political scientists, considerably less research has occurred on monitoring non-US political blogs (one exception is available elsewhere[1]). Marti Hearst, Matthew Hurst, and Susan Dumais note that it's difficult to "take the pulse of the populace" on a given topic and to identify which blogs are worth reading on that topic.[2] The project we discuss in the main text tries to address these issues. Rebecca Goolsby has detailed several "illusions" and "delusions" to be avoided in analyzing foreign blogs for intelligence.[3] Researchers at Harvard's Internet & Democracy project present a clustering algorithm to discover affinities among Persian-language blogs based on the URLs to which they link, revealing interesting substructures.[4] Finally, researchers on IBM's Banter project present a sophisticated model of political blog topics.[5]

References

1. H.W. Park and M. Thelwall, "Developing Network Indicators for Ideological Landscapes from the Political Blogosphere in South Korea," *J. Computer-Mediated Communication*, vol. 13, no. 4, 2007, pp. 856–879.
2. M.A. Hearst, M. Hurst, and S.T. Dumais, "What Should Blog Search Look Like?" *Proc. 2008 ACM Workshop on Search in Social Media*, I. Soboroff et al., eds., ACM Press, 2008, pp. 95–98; http://dx.doi.org/10.1145/1458583.1458599.
3. R. Goolsby, "The DoD Encounters the Blogosphere," *1st Int'l Conf. Social Computing, Behavioral Modeling and Prediction*, presentation slides, 2008; www.public.asu.edu/~huanliu/sbp08/Presentations/Invited/02_Goolsby_Blogosphere%20Phoenix.ppt.
4. J. Kelly and B. Etling, *Mapping Iran's Online Public: Politics and Culture in the Persian Blogosphere*, research publication no. 2008-01, Berkman Center for Internet & Society Publication Series, 2008.
5. W. Gryc et al., "Mining Political Blog Networks," *Harvard Networks in Political Science Conf.*, presentation slides, 2008; www.hks.harvard.edu/netgov/files/NIPS/gryc_politicalblog_networks_NIPS2008.ppt.



(a) Malaysian Sopo blogosphere (b) US political blogosphere (Adamic & Glance, 2005[4])

Figure 1. Link structures in the blogosphere. We compare the (a) Malaysian social-political and (b) US blogospheres and see little evidence of polarization in the Malaysian structure.

Table 1. Profile information disclosed in social-political vs. random Malaysian blogs.

	Malaysian sopo bloggers (*N* = 700)	**Random Malaysian bloggers (*N* = 112)**
Student (%)	5.9	26.7
Gender (%)	26.5 female 48.4 male 25 unspecified	57.1 female 25.9 male 17 unspecified
Plausible real name (%)	35	26.8
Age	31.9 (average) 51.5% specify	20.5 (average) 66.9% specify
Email address (%)	48.7	37.5
Profile picture (%)	39.1	71.4
Institutional affiliation (%)	19.1	6.3
Political party (%)	27.9% specify 10.2 UMNO 7.1 PAS 4.2 PKR 2.5 DAP 1.2 PSM	5% specify
Average inbound links	53.7	2.5
Average inbound blogs	23.6	1.8
Average total comments	43.1	3.5
Average comments per commented blog post	3.5	1.3
Average PageRank	2.8	1.2

The Malaysian sopo blogs in our corpus are approximately four times more likely to be in English than Malay (Bahasa Melayu), based on language identification applied to their profile pages. (We haven't tried to deal with the small amount of Chinese blog content in our data.) The analysis we describe does include both Malaysian English and Bahasa Melayu content.

By mining automatically identified profile pages for each blog (see Table 1), we see that the Malaysian sopo bloggers were older, on average, than a random sample of Malaysian bloggers. Gender distribution was essentially reversed: sopo bloggers were about twice as likely to be male as female; the reverse is true for random Malaysian bloggers. Sopo bloggers were more likely to provide an email address, but random bloggers were more likely to provide a profile picture. Sopo bloggers were somewhat more likely to blog under a plausible real name and provide an affiliation than random bloggers, and sopo bloggers were more likely to mention a political party in their profile; their distribution follows the distribution of political affiliations in Malaysia generally. Finally, sopo bloggers were much less likely to be students. The remaining rows show that the average sopo blogger has more in-links and comments than random Malaysian bloggers.

Blogs and Blog Metrics

Contemporary information retrieval approaches model a Web page in two ways. First, search engines compute a representation of what the Web page is about (its *aboutness*). Second, search engines calculate a metric of the document's quality. In response to a query, search engines rank and return documents according

Table 2. Social media participation by activity and country.*

	US (%)	UK (%)	France (%)	Germany (%)	Japan (%)	South Korea (%)
Read blogs	25	10	21	10	52	31
Comment on blogs	14	4	10	4	20	21
Write blogs	11	3	7	2	12	18
Visit social network sites	25	21	3	10	20	35
Watch user-generated video	29	17	15	16	20	5
Upload video	8	4	2	2	3	4

*Data provided by Forrester Research and based on users who participate at least monthly. The table is adapted from C. Li and J. Bernoff, *Groundswell: Winning in a World Transformed by Social Technologies*, Harvard Business Press, 2008.

to both the degree that they're about the query terms and the degree to which they're high-quality documents.

In contemporary search engines, the aboutness computation is usually some variant of weighting terms in the document via *tf***idf* (term frequency by inverse document frequency). The search engine assigns each term a weight proportional to how frequently it appears (*tf*) in the given document and inversely proportional to how frequently it appears in other documents (*idf*). Search engines use inverse document frequency because if a term appears in many documents, it's unlikely to distinguish what the document is about. Terms might also be *boosted* or *discounted* because of where they appear in a document: terms appearing in the title, for example, are boosted, whereas those appearing in a navigational menu might be discounted. A document's meaning and, by extension, a document collection's meaning, is thus a weighted vector of the terms it contains.

Search engines (including the blog search engine Technorati) usually consider only the terms in the document itself when representing the document's topic. Google also includes incoming link text. To our knowledge, no contemporary search engine includes the text of any cited document in its representation of a document's topic.

A document's quality metric depends on the degree to which other documents link to it. Google's PageRank metric,[5] for example, measures each document's *eigenvector centrality*. That is, the PageRank algorithm calculates a document's quality recursively by counting *in-links* (hyperlinks pointing to a document) and weighting them more highly if they're from high-quality documents (those that are themselves in-linked by high-quality documents). Technorati's Authority metric for blogs and posts measures blogger *in-degree centrality*, determined by simply counting the number of inbound links from other blogs during the preceding six months.[6]

Measuring Credibility

Our blogger quality metric, which we call the credibility metric, combines three measures:

- blogger authority,
- reader engagement with the blog, and
- blogger accountability.

Using Technorati's API, we measure authority as blogger in-degree centrality, an effective way to determine the most influential blogs among other bloggers.

Although knowing what other bloggers are interested in is important, this might not be the best metric to determine general reader engagement. As Table 2 shows, more people worldwide read and comment on blogs than write them, so counting only blogger in-links potentially underestimates a blog's reception. Blog readership is difficult, if not impossible, to determine externally. So, we measure engagement with blogs through recent comment counts. Our engagement metric doesn't distinguish positive from negative comments. We believe that the ability to attract any commentary is more indicative of engagement than comment polarity; most blogs attract few readers or commenters, and, anecdotally, we've seen no blogs that attract only negative comments.

Surprisingly, comments and in-linking aren't strictly correlated, at least not in the Malaysian data we examined. We compared a log of the total in-links provided via the Technorati API and a log of the total comment counts for each

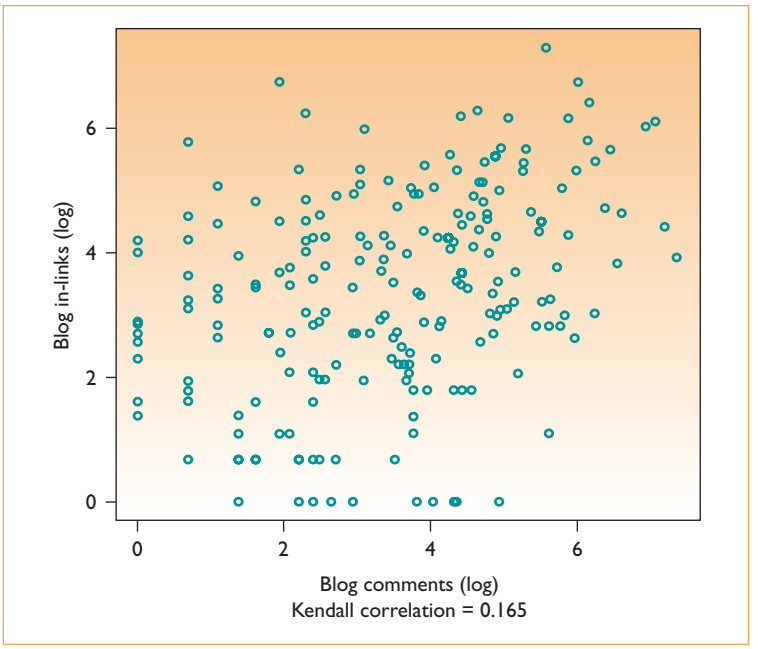


Figure 2. Blog in-links and comments. We measured the correlation between in-links and the number of comments for Malaysian sopo blogs and determined that the correlation isn't very strong.

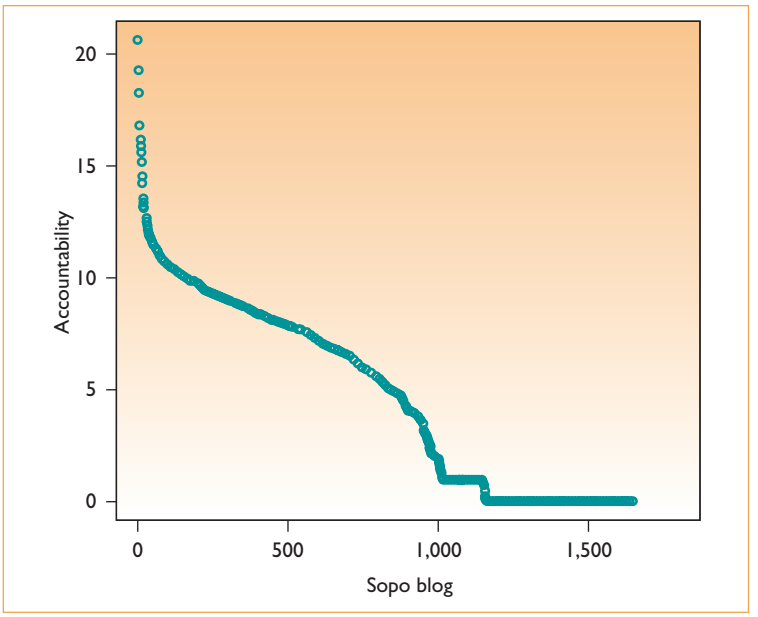


Figure 3. Unnormalized accountability scores. The average score for social-political (sopo) bloggers is 7.0, compared to 6.2 for randomly selected Malaysian bloggers.

blog crawled at depth two, which includes all the posts linked directly from the blog homepage. The correlation between blog counts and in-link counts (see Figure 2) wasn't very strong (a 0.165 Kendall correlation). Some blogs receiving many in-links receive few comments, and, more importantly, some blogs that receive many comments receive few in-links.

Finally, bloggers are accountable to the extent that others can hold them responsible for what they say. Malaysian bloggers have been held legally liable for their posts and even for comments others have left. However, anyone who risks social capital (such as reputation or relationships) by making themselves accountable for what they say is more credible than an anonymous writer who can avoid consequences for publishing his or her views. This presupposes that we can assign praise or blame to a blogger and therefore requires knowing something about his or her identity. Total anonymity is the opposite of accountability.

We compute our accountability metric as follows. For each blog profile, if a blogger lists a country, city or region, industry or occupation, email address, or Web page or IM handle, this listing contributes one point each to the blogger's accountability score. If the user lists a full name, which we identify as a multiword phrase containing no English words after eliminating titles such as "Dr." or "Auntie," then we count it as plausibly real and add one point to the aggregate accountability score. We consider every additional proper name (nonsentence initial capitalized phrase) in the free-text portion of the profile to be an affiliation, which contributes 0.5 points to the total. Because they consist of just free text, as opposed to the structured data in Blogger profiles, Wordpress profiles often contain a much longer personal information section, so we limited the number of identified affiliations to a maximum of 10. Blogging longevity is also an indicator of accountability, so we added a point for every six months a blogger has been blogging. This is computed either via the start date (Blogger) or earliest blog archive date (Wordpress).

Figure 3 shows the distribution of unnormalized Malaysian sopo blog accountability scores. The average score for sopo bloggers is 7.0 compared with an average score of 6.2 for randomly selected Malaysian bloggers.

To combine blogger authority (in-link counts), engagement (comment counts), and accountability (nonanonymity features) into a single credibility metric, we took each blog in-link count (in-degree centrality) percentile and added it to each blog's comment count percentile

Table 3. Blog-quality metrics applied to a list of influential Malaysian sopo bloggers (listed alphabetically).

Blog	PageRank percentile	Technorati Authority percentile	Comments (engagement) percentile	Reception (authority + comments) percentile	Credibility (reception + accountability) percentile
blog.limkitsiang.com	0.959	0.994	0.987	0.999	0.999
cetusan-hati.blogspot.com	0.783	0.943	0.711	0.909	0.971
chegubard.blogspot.com	0.783	0.958	0.817	0.951	0.988
drecheah.com	0	0.479	0.401	0.52	0.343
educationmalaysia.blogspot.com	0.959	0.987	0.959	0.993	0.998
elizabethwong.wordpress.com	0.959	0.932	0.903	0.97	0.994
harismibrahim.wordpress.com	0.959	0.978	0.989	0.996	0.999
jeffooi.com	1	1	0.873	0.98	0.996
khookaypeng.blogspot.com	0.959	0.919	0.705	0.891	0.963
kickdefella.wordpress.com	0.959	0.984	0.613	0.88	0.966
niamah.blogspot.com	0.783	0.996	0.964	0.996	0.996
niknazmi.com	0.783	0.974	0.72	0.924	0.975
othermalaysia.org	0.395	0	0.904	0.542	0.877
rantingsbymm.blogspot.com	0.959	1	0.94	0.991	0.977
rockybru.blogspot.com	0.959	0.89	0.761	0.909	0.771
sloone.wordpress.com	0.959	0.978	0.74	0.935	0.980
teresakok.com	0.783	0.994	0.672	0.915	0.973
tianchua.net	0.783	0.978	0.74	0.935	0.972
Average	**0.744**	**0.888**	**0.800**	**0.967**	**0.930**

*Influential blogs were identified in Jun-E Tan's and Zawawi Ibrahim's study.[7]

(both normalized 0.0 to 1.0) to get a combined *reception* score. We then added the accountability measure percentile, also normalized from 0.0 to 1.0. Because the Malaysian sopo blogosphere is relatively small, we can compute this score for each blog directly (as opposed to PageRank, which is usually recursively estimated, rather than computed directly).

To validate our approach to blogger credibility, we calculated the reception (in-links and comment counts only) and credibility (reception plus accountability) score for every Malaysian sopo blog as described, then compared our metrics with other blog-quality metrics (see Table 3). After ranking all 1,272 blogs in our corpus, we compared rankings produced by the most influential Malaysian blogs' various metrics, as determined independently by political scientists Jun-E Tan and Zawawi Ibrahim in their monograph on Malaysian bloggers' impact on the 2008 election.[7] (Of the 25 blogs Tan and Ibrahim mentioned, two weren't included in our comparison because they had been discontinued at an indeterminate time: malaysiavotes.org and hishamspeaks.wordpress.com. We counted Raja Petra Kamarudin's *Malaysia Today* site [www.m2day.org] as a news portal, not a blog, because it aggregates news from other sites and provides original reporting from numerous contributors.) We compared how highly these independently selected blogs were ranked according to various metrics. Tan and Ibrahim didn't themselves rank the bloggers they specified as most influential, so

Table 4. Top 20 Malaysian sopo bloggers by credibility score (1,272 total blogs).

Rank	Blogger and blog URL	Party affiliation	Credibility score
1	Lim Kit Siang, member of parliament; http://blog.limkitsiang.com/	DAP	2.974
2*	Haris Ibrahim, lawyer; http://harismibrahim.wordpress.com/		2.96
3*	Tony Pua, member of parliament; http://educationmalaysia.blogspot.com	DAP	2.933
4	Capt. Yusof Ahmad, maritime arbitrator; http://cyusof.blogspot.com		2.858
5*	Jeff Ooi, member of parliament; www.jeffooi.com	DAP	2.856
6	Patrick Teoh, actor/radio personality; http://niamah.blogspot.com		2.854
7*	Elizabeth Wong, Bukit Lanjan assemblyperson ; http://elizabethwong.wordpress.com	PKR	2.829
8	"Hamka," blogs about Islam; http://keretamayat.blogspot.com		2.829
9	"aNIe," self-described housewife; http://kaklady.blogspot.com		2.819
10	"Maverick SM," construction management; http://maverickysm.blogspot.com		2.816
11	Azly Rahman, political columnist/scholar living in US; http://azlyrahman-illuminations.blogspot.com		2.782
12	"Ashterix", IT blogger; http://ashterix.blogspot.com		2.781
13	Ahmad Atalib, journalist; http://ahmadatalib.blogspot.com		2.770
14†	Badrul Hisham Shaharin; http://chegubard.blogspot.com	PAS	2.765
15	Zainul Abidin, retired naval commander; http://zainulabs.blogspot.com		2.755
16	Zewt; http://zewt.blogspot.com		2.736
17	Husin Lempoyang; http://the-antics-of-husin-lempoyang.blogspot.com		2.734
18	Johnny Ong, construction industry; http://johnny-ong.blogspot.com		2.707
19	Siasah Daily, editors of PAS-related newspaper; http://siasahdaily.blogspot.com	PAS	2.706
20	Susan Loone,; http://sloone.wordpress.com/		2.702

* indicates that the blogger won their election in 2008; †indicates an electoral loss in 2008; blogs in blue are among Tam and Ibrahim's[7] top Malaysian social-political bloggers list.

we listed them alphabetically. We compared Google PageRank metrics; blogger in-links (Technorati Authority); comment counts; comment counts and in-links combined (reception); and in-links, comments, and accountability measures combined (credibility). When comparing the blogs' average percentile for each metric, our metrics (the last two columns in Table 3) showed a significant improvement over the PageRank, Technorati Authority, and comment count metrics alone. This demonstrates that, for blogs that subject-matter experts independently identified as high quality, our measures rank these blogs higher, on average, than the other metrics.

The data in Table 3 implies that the PageRank metric (via the API at www.exslim.net/api) varies considerably among Tan and Ibrahim's top bloggers. Even the top PageRank blogger, Jeff Ooi (www.jeffooi.com), receives a raw PageRank of only 6 out of a possible 9. One influential blogger, according to Tan and Ibrahim, gets no PageRank at all.

Using Technorati's Authority metric, we see less variance among the top bloggers Tan and Ibrahim identified, although this time a different blog receives a 0 score. On average, Tan and Ibrahim's top blogs rank higher via the Authority metric than the PageRank metric.

Next, we see that if we consider only comment counts, then Tan and Ibrahim's top blogs do rather worse than if we consider only Technorati Authority. By combining our authority and engagement metrics into a single blog reception measure, we get a ranking that does the best overall at ranking the top blogs from Tan and Ibrahim.

We've adopted the combined credibility metric as our preferred metric even though it does slightly worse than the authority- and engagement-based metrics for the top bloggers. The accountability part of the credibility metric lets us distinguish new blogs with reasonable accountability scores from completely anonymous blogs.

Table 4 shows the top 20 blogs out of the

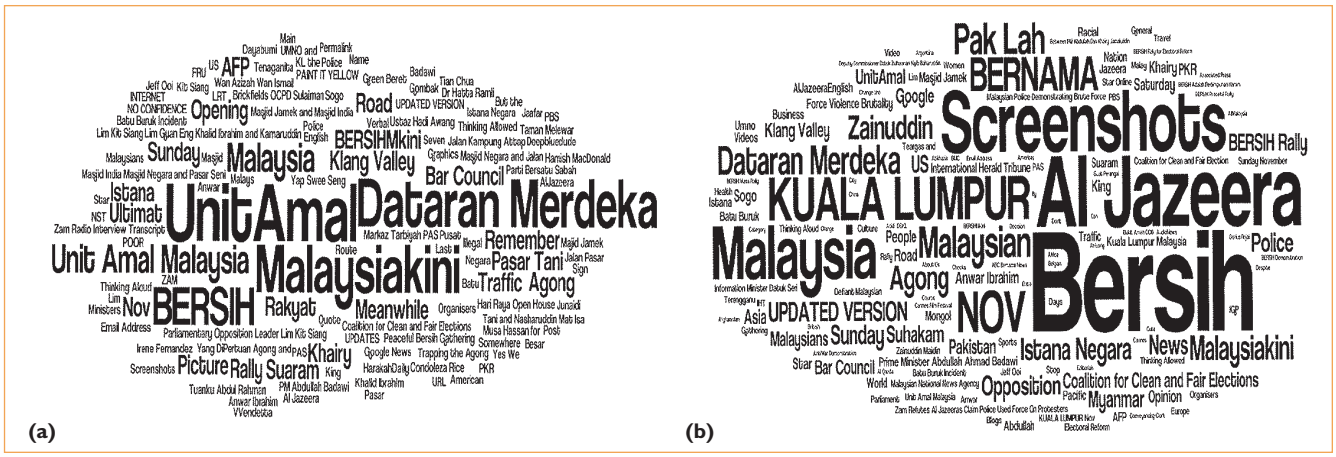


Figure 4. Term cloud visualization. We compare term frequencies in a Jeff Ooi blog post between (a) only the post's text and (b) the text as well as news stories it cites. Font size is proportional to term frequency.

1,272 total blogs we ranked, as determined by the combined credibility metric. Starred entries represent bloggers who won an election in 2008. Badrul Hisham Shaharin ("Chegubard"), marked †, was the only blogger in the top 20 to lose his election (against Khairy Jamaluddin, the Malaysian prime minister's son-in-law). Eight of our top 20 blogs were also on Tan and Ibrahim's list.

We computed our credibility and reception metrics offline and can recompute them at regular intervals. With these metrics, we believe we've identified a scalable and practical method for assigning importance to sopo bloggers that better predicts expert judgment than competing available metrics.

Measuring Relevance

As mentioned, contemporary information retrieval algorithms typically measure what a document is about using a vector of its terms, weighted by *tf*idf*. We've rejected applying this approach directly to blog post aboutness, first because of uncertainty as to the relevant set of documents for inverse document frequency and second because not all the terms that determine blog post aboutness appear in the post's text.

Inverse document frequency limits the impact of very common words in retrieval queries,[8] but, because most topical blog queries tend to be proper nouns,[9] it seems sensible to boost the standard *tf*idf* score for proper nouns by adding a term-frequency factor. So, we index a special aboutness field, in addition to indexing the content regularly for scoring via *tf*idf*. This field contains a term-frequency-weighted vector of proper names found in a blog post or in a news story that has been cited.

To a great extent, sopo blog posts are commentaries about news stories or other items on the Web to which the blog posts link, so we consider a blog post to be about a combination of what it says and what any news story that it cites says. A blog post that's just a hyperlink to a news article (for example, "Read this!") is entirely about what it links to, not just the link text alone. A similar approach is useful for treating quotations, which are used extensively in sopo blog posts.

In Figure 4, we present term cloud visualizations (using Wordle.net) of proper names' relative frequency, as extracted by our system. Font size is proportional to term frequency. Figure 4a depicts the relative frequency of proper names in the text of just one blog post by Jeff Ooi about the Bersih rally on his "Screenshots" blog (www.jeffooi.com/2007/11/how_they_painted_it_yellow.php). Figure 4b depicts the relative frequency of proper names in the combined text of Ooi's blog post and all the news articles it cites. In the combined term cloud, the most prominent term is "Bersih," and the rest of the prominent terms indicate key players, locations, and news organizations that covered the event. (The name "Zainudin" refers to the Malaysian Information Minister, who was widely disparaged for a telephone interview with Al Jazeera TV in which he denied that the police were violently suppressing the rally.) In contrast, the most prominent terms in the post-only term cloud are "Unit Amal" (a paramilitary organization of the opposition Islamist PAS party used to keep order), "Dataran Merdeka" (the march's starting point), "Malaysiakini" (an independent news site), and, less prominently, "Bersih," the coalition that spon-

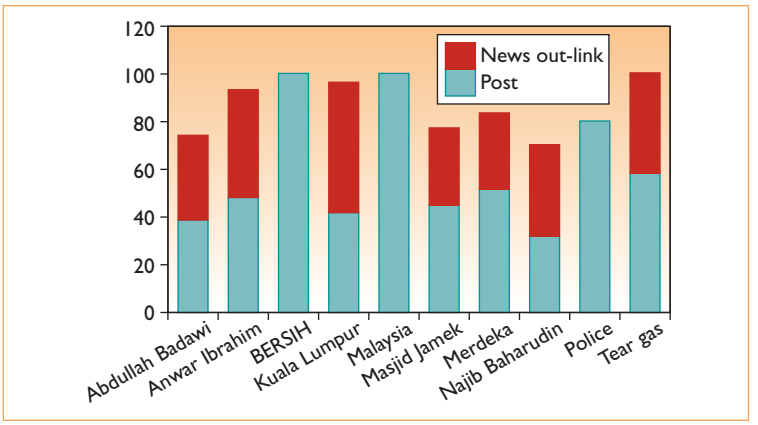


Figure 5. Term distribution in posts and out-links (% of documents containing the term). The chart shows that combining the text of posts and news out-links allows the system to better characterize and cluster posts by topic.

sored the rally. The post-and-citation term cloud clearly provides a better depiction of the post's topic. Additionally, the richer representation of post topics that this combined text-and-citation approach provides facilitates more descriptive blog post clustering by topic.

Figure 5 shows the percentage of a set of 31 posts from our Bersih rally corpus that cite one or more of 10 popular news articles about the rally. The blue portion represents the percentage of documents containing each term if we use only the blog post's text. The full column shows the percentage of documents containing each term if we also evaluate cited news articles' text. Thus, whereas 48 percent of the blog posts alone mention opposition leader Anwar Ibrahim, 97 percent of the combined texts mention this significant participant. Including this term in a cluster label would better indicate the cluster's content.

We score document clusters using the number of terms involved and the number of documents clustered, favoring more terms (because it's more descriptive) when a tie occurs. If we consider the combined post text and outlink document text, we can cluster 16 of the 31 blog posts under eight terms common to them all: police, Bersih, Masjid Jamek, tear gas, Kuala Lumpur, Abdullah Badawi, Anwar Ibrahim, and Malaysia. Alternatively, using only the terms from the post itself, the largest cluster out of 31 documents would contain 14 documents and be labeled with just three terms: police, Malaysia, and Bersih. Thus, the largest cluster with posts and citations is three times better (16 × 8 versus 14 × 3) than the largest cluster using only the post text.

Contemporary information retrieval methods don't consider hyperlinked documents' text as relevant to what a document is about. At most, they consider the link text of pages that link to a document (as with Google). We think our approach provides a better representation of blog post aboutness and allows for more descriptive clustering.

Measuring Specificity

Information retrieval specialists originally viewed inverse document frequency as a measure of term specificity.[8] However, because a local sopo blogosphere is likely to concentrate on certain political figures repeatedly, it wouldn't make sense to count a post as losing specificity every time other blog posts mention the same person or other named entity. So, we adopted a simple measure of a text's specificity as the number of unique named individuals in it. For example, "Thomas Jefferson wrote the Declaration of Independence" is intuitively more specific than, "Some person wrote some document." Extending this simple specificity metric as unique named entity counts to blog posts is straightforward. We measure a blog post's specificity as the number of unique named entities in the combined text of the blog post and any news story it cites.

In Table 5, we compare two blog posts of nearly identical length. The first contains nine news out-links concerning the Bersih rally, and the second is from essentially an online diary and contains no news out-links. Our specificity metric — the number of unique capitalized phrases (proper nouns) in the combined post and out-link texts — reflects the first post's intuitively higher specificity.

This metric makes the specificity of a blog post that quotes an entire news article the same as a post that simply cites the same news article. Furthermore, it makes the specificity of a blog post that cites a news article and partially quotes it the same as the specificity of the news article itself, if the blog post adds no new named entities. Every additional named entity not shared with a cited news article adds to the post's specificity. Finally, we can compute specificity the same way for both English and Bahasa Melayu content or any language with similar capitalization.

Measuring Timeliness

Timely engagement with current events as they

Table 5. Specificity comparison of two blog posts.

	Blog post 1*	Blog post 2†
File size	16.05 Kbytes	60.0 Kbytes
Word count	1,029 words	1,016 words
News out-link count	9	0
Specificity, post only	55 unique capitalized phrases	55 unique capitalized phrases
Specificity, post + news out-links	**563 unique capitalized phrases**	**55 unique capitalized phrases**

*Post 1: http://ninashah.vox.com/library/post/in-lolcat-it-translates-to-icanhazdemocracynao.html
†Post 2: http://blogs.myspace.com/index.cfm?fuseaction=blog.view&friendID=449654&blogID=353337981

occur is a measure of quality among sopo bloggers. Indeed, half the links that a news article receives are established within 36 hours of its publication.[10] Malaysian sopo bloggers are definitely more engaged with the news than randomly selected Malaysian bloggers. Our research shows that Malaysian sopo bloggers link to news articles about five times more often than randomly selected Malaysian bloggers, but — perhaps because of governmental press control — Malaysian sopo bloggers link to news articles only about 15 percent of the time (compared with 50 percent of US, A-list political bloggers[4]).

In our system, we compute a blog post's timeliness as the difference between its publication date and the publication date of a news source article it references. If a post cites more than one news article, we take the minimum number of days elapsed. Thus, if a post cites multiple news articles, its timeliness is the difference between the most recent article and the post-publication time. This lets us track a blog post's temporal proximity to an event that it cites, a better representation of its timeliness than its recency (see Figure 6).

We've implemented pattern-matching techniques to identify publication dates on blogs via distinctive URL formats produced by the Blogger and Wordpress platforms. This facilitates querying posts by date range. We can't rely on the last-modified HTTP header for blog post publication dates because a post with comments or a trackback will reflect that modification time. For news source out-links, if we can't completely identify a date in the news article's URL, we identify the first full date in the text based on the assumption that datelines are usually placed prominently in news stories.

We can currently extract publication dates for Wordpress and Blogger blogs automatically with 91.6 percent accuracy (based on a sample of 60 randomly selected posts). We can correctly identify cited news articles' publication date with 87 percent accuracy (based on a sample of 57 randomly selected news citations). Thus, we can only expect roughly 79 percent accuracy in the timeliness scores computed at present.

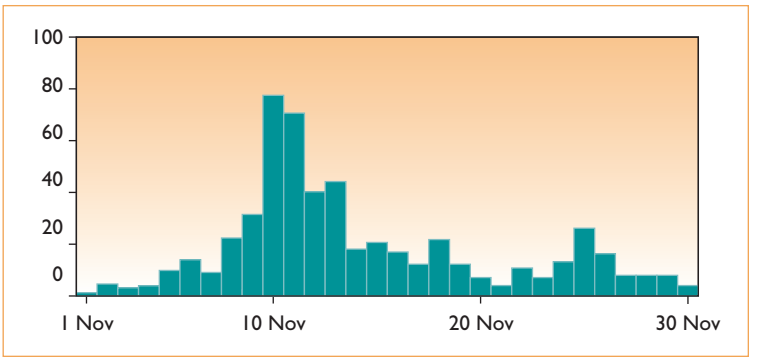


Figure 6. Temporal distribution of posts mentioning "Bersih." The Bersih rally occurred on 10 November 2007. Posts published closer to this date are timelier than those published later. Mentions of the Bersih rally in the context of a second, smaller rally that occurred in late November account for the smaller, second spike. (Data from Technorati.com.)

We can aggregate our timeliness metric across all posts in a blog to provide a metric of blog timeliness as a whole, a dimension of blog quality orthogonal to the specificity and credibility metrics. A blogger with a high degree of timeliness represents someone who is deeply committed to reacting to events as they happen.

Our metrics for representing a blog post's (and an entire blog's, by aggregation) relevance (aboutness), credibility, timeliness, and specificity are scalable and practical in that they rely on little more than identifying proper names and other features in blog posts that the system can identify automatically with high accuracy. These features include in-link and out-link URLs determining the blog's link structure, dates, proper nouns (capitalized phrases), comments, and blog

profile features, such as place names, occupations, titles, and so on. We believe that our metrics are applicable to other sopo blogospheres as well as that of Malaysia. In future work, we plan to validate our metrics with respect to other national or regional sopo blogospheres.

Acknowledgments

This material is based on work supported by the US Air Force Office of Scientific Research under contract number FA9550-06-C-0023. The views and findings expressed here don't necessarily reflect the views of the US Air Force. We also acknowledge the Nutch, igraph, and dbPedia open source communities for their efforts.

References

1. *Communications and Multimedia Act 1998*, act 588, part 1, preliminary section 3, www.skmm.gov.my/the_law/ViewAct_Part_Chapter_Section.asp?cc=45434214&lg=e&peb=n&tarid=900722&ta_prid=579037&ta_p_crid=221539&ta_p_c_srid=491924.
2. *Annual Report — Malaysia*, Reporters without Borders, 2009; www.rsf.org/article.php3?id_article=25659.
3. "Malaysia Blocks Anti-Government News Web Site," Associated Press, 28 Aug. 2008, www.usatoday.com/money/topstories/2008-08-28-2613151863_x.htm.
4. L.A. Adamic and N. Glance, "The Political Blogosphere and the 2004 US Election: Divided They Blog," *Proc. 3rd Int'l Workshop Link Discovery* (LinkKDD 05), ACM Press, 2005, pp. 36–43; http://dx.doi.org/10.1145/1134271.1134277.
5. L. Page et al., *The PageRank Citation Ranking: Bringing Order to the Web*, tech. report, Stanford Digital Library Technologies Project, 1998; http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.31.1768.
6. D. Sifry, "State of the Live Web," Apr. 2007; www.sifry.com/alerts/archives/000493.html.
7. J.-E. Tan and Z. Ibrahim, *Blogging and Democratization in Malaysia: A New Civil Society in the Making*, Strategic Information and Research Development Center (SIRD), 2008.
8. K. Spärck Jones, "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *J. Documentation*, vol. 28, no. 1, 1972, pp. 11–20; http://portal.acm.org/citation.cfm?id=106782.
9. G. Mishne and M. de Rijke, "A Study of Blog Search," *Proc. 28th European Conf. IR Research* (ECIR 06), LNCS: Advances in Information Retrieval, Springer, 2006, pp. 289–301.
10. Z. Dezso et al., "Dynamics of Information Access on the Web," *Physical Review*, vol. E 73, no. 6, art. no. 066132, part 2, 2006.

Brian Ulicny is a senior scientist at VIStology in Framingham, Massachusetts. His current research interests include the application of information retrieval and computational linguistics to Semantic Web technologies and social media analytics. Ulicny has a PhD in linguistics and philosophy from MIT. He's a member of the Gesellschaft für Semantik. Contact him at bulicny@vistology.com.

Christopher J. Matheus is the chief technology officer at VIStology in Framingham, Massachusetts. His current research interests include the development of formal ontologies and automated reasoning techniques and their application toward the intelligent analysis of heterogeneous data sources. Matheus has a PhD in computer science from the University of Illinois at Urbana-Champaign. He was a university fellow and cognitive science/AI fellow at UIUC, is a Leslie Warner Technical Achievement Award winner, and is a former T.J. Watson fellow. Contact him at cmatheus@vistology.com.

Mieczyslaw "Mitch" M. Kokar is a professor in the Department of Electrical and Computer Engineering, Northeastern University and the president of VIStology in Framingham, Massachusetts. He's an expert in information fusion, ontology-based modeling, formal reasoning, software engineering, and the application of these disciplines to the Semantic Web, situation awareness, and adaptable software. Kokar is a member of the editorial board of the *Information Fusion* journal, a senior member of the IEEE, and a member of the ACM. Contact him at mkokar@ece.neu.edu.

cn *Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.*