Newspapers vs. Blogs: Who Gets the Scoop?

Levon Lloyd, Prachi Kaulgud, Steven Skiena
Department of Computer Science
State University of New York at Stony Brook
Stony Brook, NY 11794-4400
{lloyd, prachik, skiena}@cs.sunysb.edu

Abstract

Blogs and formal news sources both monitor the events of the day, but with substantially different frames of reference. In this paper, we report on experiments comparing over 500,000 blog postings with the contents of 66 daily newspapers over the same six week period. We compare the prevalence of popular topics in the blogspace and news, and in particular analyze lead/lag relationships in frequency time series of 197 entities in the two corpora. The correlation between news and blog references proved substantially higher when adjusting for lead/lag shifts, although the direction of these shifts varied for different entities.

Introduction

Blogs represent an interesting new frontier for text analytics. More than just text, they provide significant structural information about the author, such as precise timestamps, geographical location, age, gender, and explicit friendship links. They also provide a forum for a much larger and potentially representative group of correspondents than conventional media. According to (Nardi, Schiano, & Gumbrecht 2004), people blog to express their opinions on issues to influence others. They provide examples of bloggers posting links to on-line publications, and then adding their personal commentary. Thus blogging analysis can be used to determine the collective public opinion on current events.

In this paper, we introduce our *Lydia* text analysis system as a tool for analyzing the blogspace. In particular, we compare the content of blogs with that of major U.S. newspapers over the same time frame. Our analysis helps to shed light on questions of whether the conventional news media leads or lags popular opinion as expressed in blogs.

In particular, we present the results of a quantitative analysis of the temporal relationship between news and blogs. How often do bloggers report a story before newspapers? And conversely, how often do bloggers react to news that has already been reported? There has been at least one well documented (Ashbee 2003) example of bloggers breaking a story and influencing the mainstream media. News media gave little initial coverage to controversial comments made by Trent Lott in December of 2002. However, these comments quickly became the focus of buzz in blogspace. This forced the mainstream media to revisit the story, which ultimately lead to the resignation of Trent Lott.

Main Contributions

- We introduce our *Lydia* system for analysis of blogs and online news sources. *Lydia* performs a variety of interesting analysis on named entities in text, breaking them down by source, location and time. Associations between entities are identified through juxtaposition analysis.

  Lydia is the foundation of our system for doing real-time analysis of news streams. We encourage the reader to visit our websites for our latest analysis of roughly 500 U.S. daily newspapers (www.textmap.org) and hundreds of thousands of blog postings (www.textblg.org).
- We discuss the modifications necessary to optimize the ability of our existing text analytics system to accurately process blogs.
- Visualizing the geographic location of the buzz surrounding a topic can provide insights into where there is interest in that topic. We present our system for generating *heatmaps* (Mehler *et al.* 2005) and show results of applying this to blogspace. In particular, we exhibit maps showing the geographic distribution of slang terms and entities of local interest.
- We contrast the relative blog/news interest among a wide range of topics. While certain differences are readily predictable (e.g. greater interest in popular culture on blogs), we were surprised by the degree of topic coherence between news and blogs.
- We study the lead-lag relationship between blogs and the news media, by analyzing the reference frequency time series of 197 entities which appeared often in both sources. Of these, 30 exhibited no lead/lag relationship, 73 had news leading the blogs, and 94 had blogs leading the news.

Related Work

Several existing web services allow users to navigate blogspace and provide aggregate information on what is happening there. Indeed:

- Google[1] has recently launched a blog only search engine.

[1] http://www.google.com/blogsearch/

- Technorati[2] uses blogger-provided annotation to provide a Yahoo! style directory of blogs. They also provide links to the most popular news stories, books, and movies mentioned in blogs and rank blogs based on the number of other blogs that link to it.
- Blogrunner[3] correlates news headlines with blog postings, providing links both to news articles and blog postings on the news item.
- Blogstreet[4] provides a few different services. Blog neighborhoods clusters blogs based on blogroll links between them. Blog influence quotient(BIQ) is their method of ranking blogs based on the link structure of blogspace. They also show the top books, DVDs, and CDs mentioned in blogs.

(Gruhl *et al.* 2005) show that increased chatter about a book in blogs can predict an increase in the Amazon sales rank for that book. They correlate a time series representing the number of blogs that mention a book's title with a time series representing the sales rank of the book and show that for a certain fraction of the books, the two time series were highly correlated, with the blog time series leading the sales rank time series more than half the time.

(Liben-Nowell *et al.* 2005) study friendship links between bloggers and their corresponding geography. They show that a large fraction of friendship links can be explained geographically. Further, they show that geographic distance is not the correct way to calculate the probability that two people are friends, but rather the number of people that live closer to one person than the other.

(Kumar *et al.* 2003; 2004) give an overview of the characteristics of blogs and bloggers. They studied the personal information provided by bloggers, including age, geographic location, and personal interests. They show correlations between friendship, location, age and shared interests. A study of blogspace on the community level is also presented. Algorithms for identifying these communities and temporal bursts of activity within these communities are given.

(Lin & Halavais 2004) study the geographic information that is attached to blogs. They present an algorithm that looks not only at the profile page of a blogger, but also at other clues such as Who-is data for self-hosted blogs and links to local weather and news present in the blog. The first three digits of the zip code of a blogger is proposed as a way to more robustly represent the location of a blogger for studying the geography of blogs.

(Gruhl *et al.* 2004) presents a study on information diffusion in blogspace along two dimensions, topical and individual. First, they present a model of day-to-day changes in the number of blog postings about a topic, showing that this is the sum of background chatter about the topic and event based spikes. Then they study the propagation of topics from blogger to blogger by comparing this propagation to the spread of an infectious disease through a population. They present a model describing the probability that a person blogs about a topic based on the probability that one of his friends blogs on that topic and then validate this model with both synthetic and real data.

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

[3]http://www.blogrunner.com/

[4]http://www.blogstreet.com/

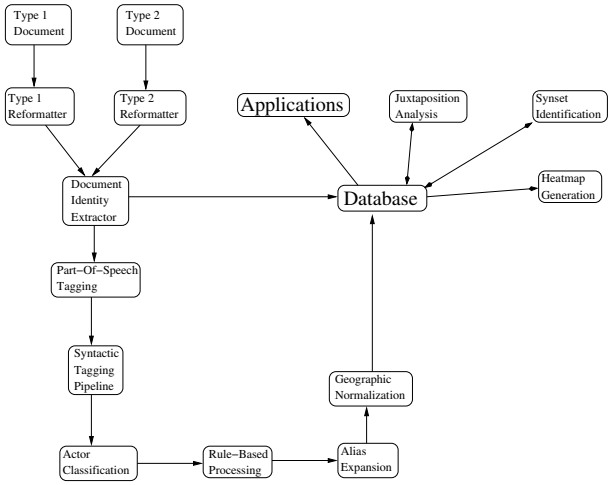

Figure 1: Block diagram of the *Lydia* Pipeline

The Newsblaster project (McKeown *et al.* 2002; Barzilay, Elhadad, & McKeown 2002) provides summaries of the day's news that it obtains from a set of online newspapers. Applying state-of-the-art techniques in topic detection and tracking, they cluster articles by event, group these clusters into groups of articles about related events, and categorize each event into pre-determined top-level categories(U.S., World, Sports, Entertainment, Science/Technology, and Finance). They have recognized different classes of articles: articles about a single event, articles about multiple events, biographical articles, and articles that do not fit into any of the other categories. Different multi-document summarizers are used, one designed for each class of article. Each cluster of articles is routed to one of these summarizers to produce the summary for that event.

Lydia

Lydia (Lloyd, Kechagias, & Skiena 2005) is a system designed for high-speed analysis of online text. We seek to analyze hundreds of text feeds daily. *Lydia* is capable of retrieving a daily newspaper like *The New York Times* and then analyzing the resulting stream of text in roughly one minute of computer time. We are capable of processing over $500,000$ blog postings per day on a single commodity computer.

A block diagram of the *Lydia* processing pipeline appears in Figure 1. The major phases of our analysis are:

- *Spidering and Article Classification* – We obtain our newspaper and blog text via spidering and parsing programs which require surprisingly little customization for different news sources. We also attempt to classify source articles by news type (e.g. business, sports, entertainment).
- *Named Entity Recognition* – Identifying where *entities* (people, places, companies, etc.) are mentioned in news-

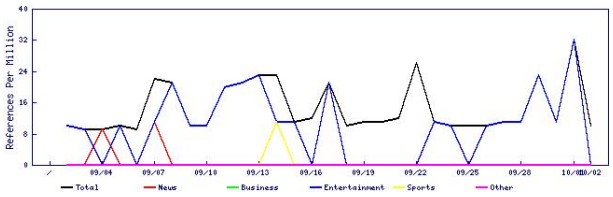

Figure 2: Blog post classifications for Jamie Foxx

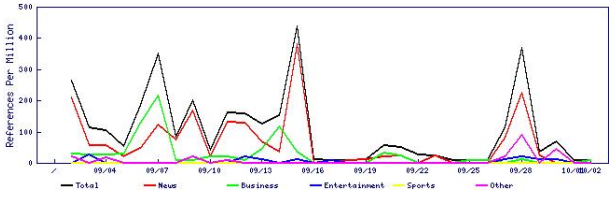

Figure 3: Blog post classifications for Michael Brown

paper articles is a critical first step in extracting information from them.

- *Juxtaposition Analysis* – For each entity, we wish to identify what other entities occur near it in an overrepresented way. We use statistical methods for ranking such co-occurrence relations.
- *Co-reference Set Identification* – A single entity is often mentioned using multiple variations on their name. For example, *George Bush* is commonly referred to as *Bush*, *President Bush* and *George W. Bush*. Further the abundance of misspellings in blogs leads to multiple variations of the same name (e.g. *Britney Spears*, *Brittany Spears*). Our methods for identifying such co-reference sets are discussed in more detail in (Lloyd, Mehler, & Skiena 2005), with a performance evaluation.
- *Temporal and Spatial Analysis* – We can establish local biases in the news by analyzing the relative frequency with which given entities are mentioned in different news sources.

Classification of News/Blog Entries

Our system for analyzing newspapers includes Bayesian classification of newspaper articles into one of *news*, *business*, *entertainment*, *sports*, and *other*. We were pleasantly surprised with how well our classifier worked on blog postings, for which a priori classifications did not exist and on

Class	% of Blog Entries	% of News Articles
business	1.44%	28.89%
entertainment	42.81%	12.31%
news	35.71%	33.97%
other	4.80%	6.52%
sports	15.23%	18.31%

Table 1: Percent of blog entries and news articles classified in each class

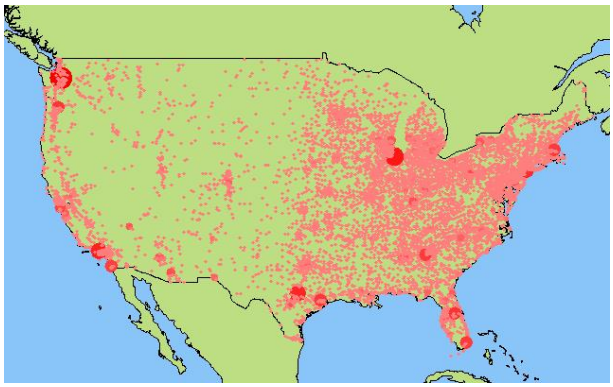

Figure 4: Map of locations of all of the bloggers in our database

which it was not trained. For example, observe how the reference frequency classifications (denoted by line colors) differ for entities such as entertainer Jamie Foxx (Figure 2) and government official Michael Brown (Figure 3).

Table 1 shows how blog postings and news articles are distributed into each of our classes. The two most interesting differences are the larger number of blog posting classified as entertainment and the small number of blog postings classified as business. This is one way of comparing the content of news to that of blogs and shows that people blog more about the entertainment world and less about the business world.

Examining our article classification more closely, we found that entertainment figures had most postings in the entertainment category, and political figures had most postings in the news category. Table 2 shows that we can categorize people based on the predicted type of the blog postings that they appear in. Entertainment figures (*Jamie Foxx*, *Jessica Alba*, ...), political figures (*Michael Brown*, *President Bush*, *Cindy Sheehan*), and sports figures (*Derek Jeter*, *Terrell Owens*) have most mentions classified in entertainment, news, and sports respectively.

Heatmaps

Visualizing the geographic location of the buzz surrounding a topic can provide insights into where there is interest in that topic. Such analysis requires a geographically diverse set of bloggers. Figure 4 gives the geographic coverage of our set of bloggers, where we drew a dot at each city that has a blogger in it scaled according to the number of local bloggers. Figure 4 shows heavy coverage of the Eastern United States but the apparently low uniform coverage in the West reflects population density. Indeed, our database has at least one blogger in $9,485$ of the roughly $25,000$ different cities in the U.S.

Our heatmaps do show interesting geographic distributions of both slang and certain entities. Figure 5 shows the heatmap for *Steak N Shake*, a fast-food restaurant with most of its locations in the Mid-West[5]. The buzz correlates well

[5]http://www.steaknshake.com/states/location.asp

Entity Name	News Articles	Business Articles	Entertainment Articles	Sports Articles	Other Articles
Michael Brown	180	81	14	0	22
President Bush	716	809	106	3	103
Cindy Sheehan	251	93	36	3	29
Jamie Foxx	5	0	39	2	0
Jessica Alba	8	0	84	5	3
Martha Stewart	33	8	190	4	6
Michael Jackson	41	3	649	15	14
Derek Jeter	7	0	3	22	0
Terrell Owens	1	0	1	17	0

Table 2: Number of blog postings in each class for a given entity.

Figure 5: Heatmap for Steak N Shake

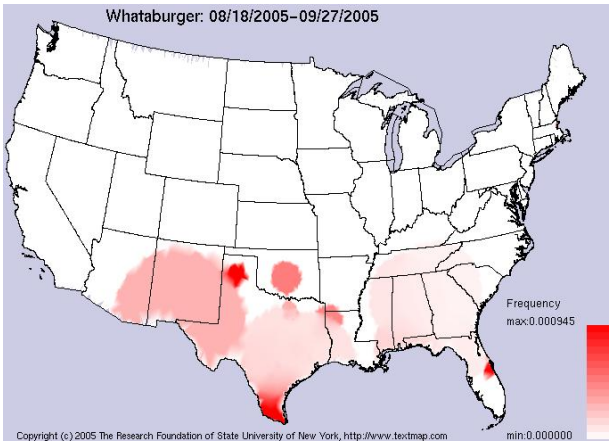

Figure 6: Heatmap for Whataburger

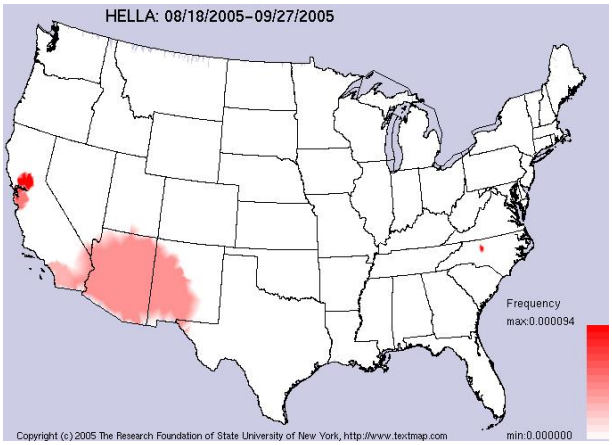

Figure 7: Heatmap for Hella

with where its locations are. The same phenomenon holds for *Whataburger*, a Texas-based restaurant chain. Figure 7 shows the heatmap of the slang term *hella*, used primarily on the West Coast. Our analysis reflects this, showing usage concentrated around *San Francisco* and *Los Angeles*.

To produce heatmaps, we extract the location of the blogger from their user-information page. For each city, we approximate a sphere of influence based on the number of bloggers there, its population, and the populations of surrounding cities. The heat an entity is generating in a city is now a function of the number of times it is mentioned in the other cities that have influence over that city.

To draw the map, we triangulate the cities in the continental U.S.. Then we determine the heat at each city. Finally, we color each triangle by interpolating the heat at each vertex. Details appear in (Mehler *et al.* 2005).

Blog Specific Processing

Blogs are more difficult to analyze using traditional natural language processing techniques than professional news reports. While newspapers adhere to standard usage of grammar, punctuation and capitalization, blogs are seldom grammatically correct, often use inconsistent capitalization and punctuation, and contain typos, misspellings, and unique abbreviations. Adapting our *Lydia* text analysis pipeline

from newspapers to blogs required a modest amount of customization, which we report on in this section.

The first adaptation necessary for blogs was to account for inconsistent capitalization. In newspaper articles, words entirely in upper case are usually acronyms. In the informal language of blogs, some write in a uniform case, while others use capitalization to express their emotions such as happiness or anger (”Im SO excited!” or ”I HATE you”) or to convey a certain tone of speech (”There was NO way I was going to the party”). We explicitly remove named entity tags from any stop-word or emotion-conveying word that was tagged because of capitalization. Many bloggers use uniform case leaving us unable to use capitalization to identify named entities. Our remedy was to use a gazetteer-based approach to identify classes of named entities (e.g. cities, companies, universities, etc).

A potential problem here is the ambiguity of some of these tokens. For example, Gary is both a popular male first name and the name of a city in Indiana. We handle this situation by tagging the entity as ambiguous, and then run a Bayesian classifier that is trained to identify the semantic class of an entity based on its context.

The second necessary change was to account for the heavy usage of emoticons[6] (:), :(, etc.) and Instant Messaging short forms[7] (“b4 = before”, “2nite = tonight”, etc.). Emoticons interfere with the part-of-speech tagger because it interprets them as weird punctuation. Further, without determining their meaning, the presence of emoticons is irrelevant. Gazetteers of commonly used emoticons were used to identify and eliminate them from the text. Similarly, Instant Messaging short forms confuse the part of speech tagger because they are not words in its lexicon. To resolve these problems, the commonly used short forms are recognized and expanded to their normal English equivalents.

One interesting direction of future work in this arena is to use both the emotion-conveying capitalization and emoticons in the development of a sentiment extraction system that can capture the strong emotions that are often communicated in blogs.

Experiments, Results, and Discussion

Here we describe experiments that we performed on news/blog data. First, we compare the most popular entities in each corpus. Then, we examine individual entities and compare how the number of references to an entity changes with time in the two corpora.

All of the experiments in this paper used two sources. For blogs, we downloaded Livejournal's[8] latest posts RSS feed of the most recent blog postings every 15 minutes between August 18, 2005 and September 27, 2005. This totals over $500,000$ blog posts. Our newspaper analysis was performed on a set of 66 of the most important daily U.S. newspapers in the same time frame.

[6]http://www.netlingo.com/smiley.cfm

[7]http://www.netlingo.com/emailsh.cfm

[8]http://www.livejournal.com/

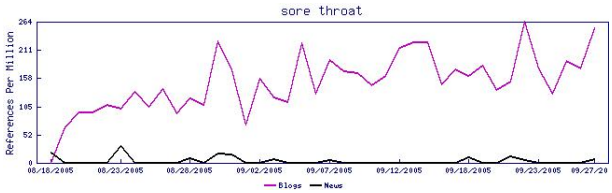

Figure 8: Time series for sore throat

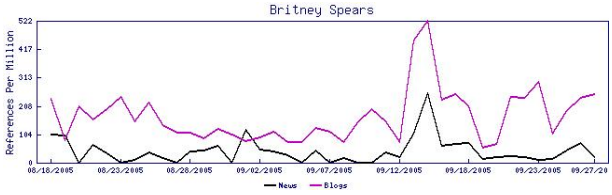

Figure 9: Blog and News time series for Britney Spears

Popular Entities in News vs. Blogs

One simple way to capture the differences in the topics that are prominent in the news and in blogs is by looking at the differences in the lists of the most popular entities that are referred to in both corpora.

Table 3 shows both the most popular people in the news and in blogs. Not surprisingly, all but one of the top people in the news are political figures. What may be surprising are the relatively low prominence of these figures in blogs. For example, *Ariel Sharon* does not appear in a single blog in our corpus! Alternatively, looking at the top people in blogs, it immediately becomes clear that bloggers are more interested in the entertainment world than current events and politics. *George Bush*, *Michael Brown*, and *Cindy Sheehan* are the only non-entertainment people on the list.

Table 4 shows the most referenced drugs. The blog list features a lot of drugs that people use on a daily basis (*Tylenol*, *Advil*, *Sudafed*) showing how people use blogs to document their daily lives. In contrast, the news list features a lot of drugs that are currently getting press like *Tamiflu*, which is being touted as the only effective medicine for a potential flu epidemic, and *Vioxx*, whose maker is currently under investigation for failing to disclose information about its safety.

The most referenced companies in both corpora are shown in Table 5. An interesting similarity between the two lists is the prominence of technology companies. Further, the appearance of grocery stores (*Kroger*, *Safeway*) reinforces the fact that bloggers blog to document their everyday lives.

Comparing Time Series

We compared the time series of reference frequency to entities over time in blogs and in the news. There are a variety of interesting phenomena reflected in these time series, such as the increase in illness (sore throats) as the weather changes from summer to fall (Figure 8). Figure 9 shows the two time series for *Britney Spears*. From this picture a few things are evident. (1) The background interest in Britney Spears is much higher in blogs than in the news. (2) The news of

Top People in News			Top People in Blogs		
Rank in News	Rank in Blogs	Name	Rank in Blogs	Rank in News	Name
1	2	George Bush	1	380	Harry Potter
2	48	John Roberts	2	1	George Bush
3	2498	Ray Nagin	3	359	Britney Spears
4	7	Michael Brown	4	177	Michael Jackson
5	765	Arnold Schwarzenegger	5	421	Tim Burton
6	2975	Steve Spurrier	6	439	Kelly Clarkson
7	324	William Rehnquist	7	4	Michael Brown
8	192	Kathleen Blanco	8	16	Cindy Sheehan
9	N/A	Ariel Sharon	9	N/A	Brad Renfro
10	109	Pat Robertson	10	1921	Rick Perry

Table 3: Top People in the News and in Blog Postings

Top Drugs in News			Top Drugs in Blogs		
Rank in News	Rank in Blogs	Name	Rank in Blogs	Rank in News	Name
1	47	Vioxx	1	6	Tylenol
2	80	Tamiflu	2	63	Advil
3	80	Zyprexa	3	63	Vicodin
4	9	Viagra	4	49	Allegra
5	53	Lipitor	5	56	Prozac
6	1	Tylenol	6	72	Benadryl
7	195	Plavix	7	28	Zoloft
8	53	OxyContin	8	23	Valium
9	88	Celebrex	9	4	Viagra
10	195	Bextra	10	24	Sudafed

Table 4: Top Drugs in the News and in Blog Postings

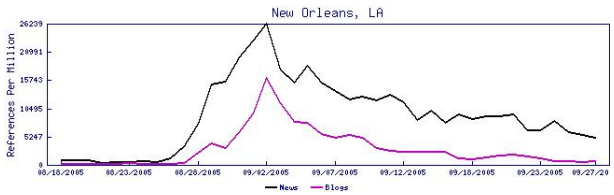

Figure 10: Blog and News time series for New Orleans

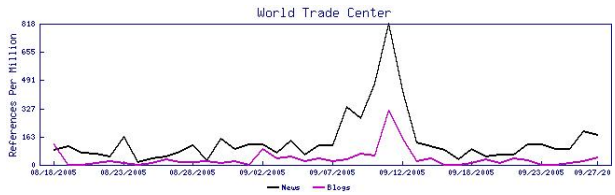

Figure 11: Blog and News time series for World Trade Center

her giving birth on September 14 caused some early buzz in blogspace on that day, but was not fully reported in the news until the next day. (3) The amount and duration of the buzz in both corpora was quite similar (relative to the amount of background noise). In the rest of this section we will discuss similar experiments to explore the relationships between the time series in the news and blogs.

We explored the correlation between news/blog reference time series, expecting high correlation. In Figure 10, we see that *New Orleans* rose to prominence during and immediately after *Hurricane Katrina* and then slowly fell back down to the background level of discussion. In Figure 11, we see how the *World Trade Center* quickly came into and then left prominence around the anniversary of the terrorist attack on it.

To study these correlation effects more rigorously, we constructed a set of 197 entities that are popular in both corpora and computed the correlation coefficient between the time series for each entity in the set. Surprisingly, the average correlation was only .085.

We noticed one interesting departure from highly correlated pictures that we were able to learn something from. Figure 12 shows the time series for *Martha Stewart*. The news time series shows three prominent spikes, while the blog time series only shows one, which is highly correlated with the last of the news time series. The first spike in the news series corresponds to reporting on a few details about one of her upcoming television shows, while the second and third spikes correspond to the premiere of her shows. It ap-

Top Companies in News			Top Companies in Blogs		
Rank in News	Rank in Blogs	Name	Rank in Blogs	Rank in News	Name
1	17	Boeing	1	2	Microsoft
2	1	Microsoft	2	4	Oracle
3	30	Merck	3	8	Home Depot
4	2	Oracle	4	40	Safeway
5	8	Walt Disney	5	23	Comcast
6	244	Chevron	6	54	Kroger
7	10	Intel	7	36	Coca-cola
8	3	Home Depot	8	5	Walt Disney
9	121	Exxon Mobil	9	13	Motorola
10	244	Northwest Airlines	10	7	Intel

Table 5: Top Companies in the News and in Blog Postings

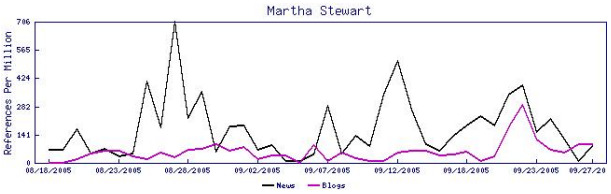

Figure 12: Blog and News time series for the Martha Stewart

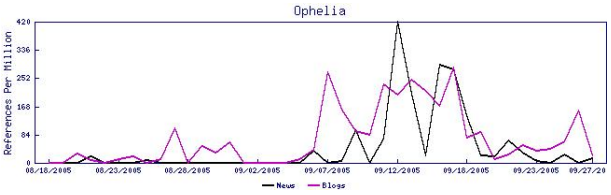

Figure 13: Blog and News time series for Ophelia

pears that the lack of a corresponding spike in the blogs for the premiere of her television show displays a lack of interest in the show from blogspace.

Perhaps this surprising lack of correlation was due to time-shifts between the two time series? Maybe blogs report things first because it is easier to post a blog entry than to publish a newspaper article, or maybe the news reports things first and bloggers subsequently comment on what they have seen in the news. Both phenomena appear to be at work. Figure 13 shows the time series for *Ophelia*, a storm that hit the east coast of the U.S. in early September.

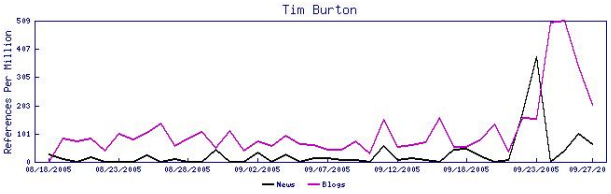

Figure 14: Blog and News time series for Tim Burton

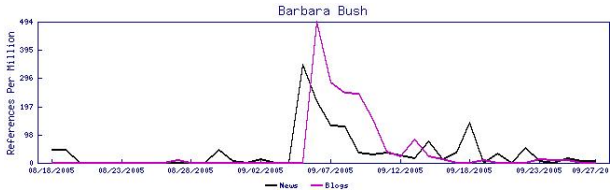

Figure 15: Blog and News time series for Barbara Bush

Correlation Coefficient	Best Lag	Name
0.990	0	Don Adams
0.898	0	Pat Robertson
0.876	1	Galveston, TX
0.857	-1	Mobile, AL
0.855	0	New Orleans, LA
0.838	-1	Barbara Bush
0.823	3	Baton Rouge, LA
0.776	1	Houston, TX
0.775	-2	Tim Burton
0.738	0	North Korea

Table 6: Entities with the highest correlations between the news and blogs

It appears that bloggers began discussing the storm about 5 days before it hit the mainstream media. On the other hand, Figures 14 and 15 show cases where bloggers lagged behind the news. The spike on September 23 for *Tim Burton* corresponds to the release of his new movie, *"Corpse Bride"*, with the corresponding spike in blogspace happening the next day and lasting for the weekend. Presumably, this is caused by people seeing the movie on the few days after its release and commenting on it in their blogs. Figure 15 shows the time series for *Barbara Bush*. The spike in the news on September 5 corresponds to reports of controversial comments she made regarding Hurricane Katrina. These comments did not cause a buzz in blogspace until the next day.

To explore this relationship statistically, we computed the

	-5	-4	-3	-2	-1	0	1	2	3	4	5
No. of Entities	20	14	14	23	23	30	12	17	13	15	16
Avg. Correlation	0.313	0.244	0.303	0.311	0.407	0.395	0.376	0.274	0.295	0.249	0.218

Table 7: Number of entities with each optimal lag and there average correlation

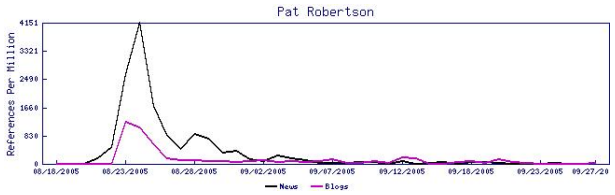

Figure 16: Blog and News time series for Pat Robertson

optimum lag and its corresponding correlation coefficient for each entity in our set. Here a positive lag means that the newspapers are leading the blogs. The resulting average optimal correlation coefficient is .317, which is much higher than with no lags. The average optimal lag is -0.213 which suggests that there are about the same number of entities for which the optimum correlation coefficient has the blogs leading as trailing.

Table 6 shows the optimum lag and corresponding correlation coefficient for the most correlated entities. From this we can see that the most correlated entities are ones which have spikes during our time period. For example, *Pat Robertson*, whose time series is shown in figure 16, got coverage for controversial comments that he made. We also can see that the optimum lag for theses entities is almost always almost 0 or 1.

Table 7 shows how many of our entities had each optimum lag and the average correlation coefficient for each of them. While most of the entities had an optimum lag of zero, the entities are well distributed among the different optimal lags. Also, while there are slightly more entities for which the blogs lead the news, there is no clear trend of blogs leading the news or news leading the blogs. A nearly equal number of entities fall on both sides.

Future Work

We leave many interesting directions of future work. First, we will explore how what is popular in the news and blogs over a larger time scale (weekly, monthly, yearly) changes. We will also try to separate background noise from discussion that is generated by outside events in blogspace and correlate these events with reporting in the news. Finally, we are building a sentiment extraction system tuned to work well on blogs, taking into account features that are unique to blogs such as emoticons and emotion-conveying capitalization.

Acknowledgements

We thank Alex Kim for his help in developing the pipeline, Manjunath Srinivasaiah for his work on making the pipeline more efficient, Andrew Mehler for his Bayesian classifier and rules processor, Izzet Zorlu for his web interface design, Namrata Godbole for her work on text markup and Michael Papile for his geographic normalization routine.

References

Ashbee, E. 2003. The Lott resignation, 'Blogging', and American Conservatism. *Political Quarterly* 74(3):361–370.

Barzilay, R.; Elhadad, N.; and McKeown, K. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artifical Intelligence Research (JAIR)* 17:35–55.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*.

Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; and Tomkins, A. 2005. The predictive power of online chatter. In *Proceedings of the Eleventh ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD)*.

Hatzivassiloglou, V.; Gravano, L.; and Maganti, A. 2000. An investigation of linguistic features and clustering algorithms for topical document clustering. In *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, 224–231.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2003. On the bursty evolution of blogspace. In *Proceedings of the Twelfth International World Wide Web Conference*.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2004. Structure and evolution of blogspace. *Communications of the ACM* 47:35–39.

Liben-Nowell, D.; Novak, J.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Science* 102:11623–11628.

Lin, J., and Halavais, A. 2004. Mapping the blogosphere in america. In *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*.

Lloyd, L.; Kechagias, D.; and Skiena, S. 2005. Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval(SPIRE 2005)*.

Lloyd, L.; Mehler, A.; and Skiena, S. 2005. Finding sets of synonymous names across documents in a large corpus. in preparation.

McKeown, K.; Barzilay, R.; Evans, D.; Hatzivassiloglou, V.; Klavans, J.; Nenkova, A.; Sable, C.; Schiffman, B.; and Sigelman, S. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of HLT 2002 Human Language Technology Conference*.

Mehler, A.; Wang, Y.; Bao, Y.; Li, X.; and Skiena, S. 2005. Heatmaps: Showing the geographic location of interest in entities in the news. in preparation.

Nardi, B.; Schiano, D.; and Gumbrecht, M. 2004. Blogging as social activity, or, would you let 900 million people read your diary? In *ACM Conference on Computer Supported Cooperative Work*, 222–231.