Access: News and Blog Analysis for the Social Sciences

Mikhail Bautin, Charles B. Ward, Akshay Patil, and Steven S. Skiena
Dept. of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
{mbautin,charles,akshay,skiena}@cs.sunysb.edu

ABSTRACT

The social sciences strive to understand the political, social, and cultural world around us, but have been impaired by limited access to the quantitative data sources enjoyed by the hard sciences. Careful analysis of Web document streams holds enormous potential to solve longstanding problems in a variety of social science disciplines through massive data analysis.

This paper introduces the TextMap Access system, which provides ready access to a wealth of interesting statistics on millions of people, places, and things across a number of interesting web corpora. Powered by a flexible and scalable distributed statistics computation framework using Hadoop, continually updated corpora include newspapers, blogs, patent records, legal documents, and scientific abstracts; well over a terabyte of raw text and growing daily. The Lydia Textmap Access system, available through `http://www.textmap.com/access`, provides instant access for students and scholars through a convenient web user-interface. We describe the architecture of the TextMap Access system, and its impact on current research in political science, sociology, and business/marketing.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Documentation, Experimentation

Keywords

News Analysis, Blog Analysis, Social Sciences, Hadoop

1. INTRODUCTION

TextMap Access provides instant access to large-scale news analysis, enabling scholars and students to identify cultural trends and interpret a wide range of social forces. Questions like:

- Do university reputations better reflect sports or academics?
- Which people and places have greater (or lesser) global significance?
- How does the editorial bias of a newspaper (conservative vs. liberal) influence its coverage of events?

can now be easily and meaningfully studied.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

These types of studies are examples of what has only become possible with our new scalable Lydia architecture, compared to the previous version of the Lydia system described in [12]. The performance improvement is approximately 20-fold. The new Lydia system can fully process our five-year archive of over 120 million U.S. news articles in less than two weeks on our 18-node Hadoop [2] cluster. The old Lydia ran on a single machine, but even if it could scale linearly, it would take 2.5 years to process the same dataset, according to the 250 articles per hour per machine performance estimate reported in [10].

Figure 1 is an example of the analysis TextMap Access makes possible, showing the extent of the historical sweep which can be explored within the corpora we provide. In this paper, we will briefly describe how the Lydia processing framework functions, the demonstrated social science applications of Lydia, and the types of statistics and corpora made publicly available to social science researchers through TextMap Access: `http://www.textmap.com/access`.

2. PROCESSING FLOW

Lydia consists of five primary components: spidering, NLP markup, sentiment analysis, entity analysis and aggregation, and access/visualization. We will discuss each of these in turn.

- *Spidering* – Lydia spiders text sources ranging from mainstream news sources to blogs on a continual, daily basis. Our largest spidered corpus consists of more than a hundred-million news articles spanning nearly five years, from over a thousand newspapers across the country and around the world. Other corpora include text derived from Lydia's own blog spiders, as well as blog text from blog-aggregation services such as Spinn3r.
- *NLP Markup* – Starting from raw unstructured text, Lydia performs a series of natural language processing tasks. Lydia begins by applying a part-of-speech tagger, the output of which is used in a number of later stages. Next, the system performs a sequence of steps designed to identify and classify named entities. Entities are classified into a range of over 100 different categories, such as *PERSON*, *COMPANY*, *ADDRESS*, or *BODY_OF_WATER*.

  With entities marked up, the system then attempts to perform pronoun resolution, local entity co-reference, and geographic normalization. Pronoun resolution simply means that, wherever possible, the system will attempt to resolve pronoun references to the indicated entity. Local entity co-reference is similar, but refers to resolving references to the same entity under different names. Later phases of the analysis engine, beyond the scope of discussion of this paper, attempt to unify names based on more than local context.

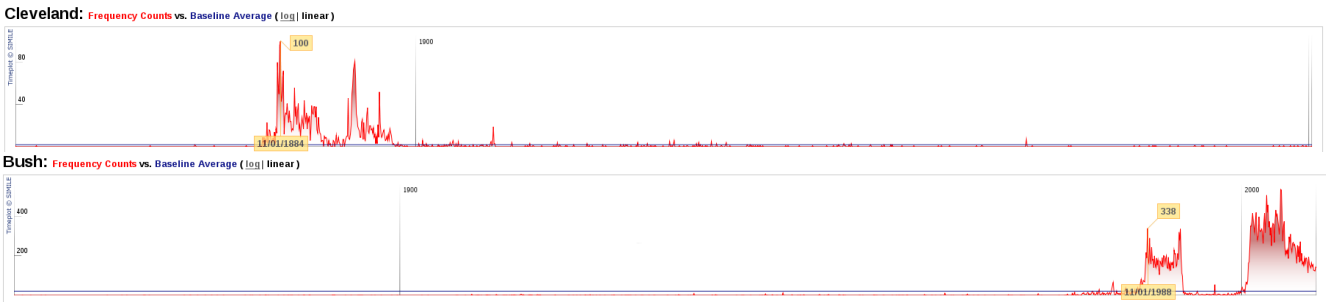

Figure 1: Frequency time series for (Grover) Cleveland and (George) Bush in the historical dataset.

- *Sentiment Analysis* – We refer the reader to the original papers [3, 8] for full details details of the Lydia sentiment analysis system, as well as Pang and Lee's excellent survey on techniques for sentiment analysis [14]. The *Lydia* sentiment analysis system is based on lexicons of positive and negative words, and associating entities with sentiment of co-occurring words from these lexicons. The *Lydia* sentiment lexicons were constructed by starting from small sets of seed words of incontrovertible polarity, targeted to each of six specific domains: *business*, *crime*, *health*, *politics*, *sports*, and *media*. The synonyms and antonyms of an electronic dictionary (Wordnet, [13]) enable us to expand each seed set into a full sentiment lexicon. Details of this process are reported in [8].

  Although validation of sentiment analysis is a difficult problem, the accuracy and usefulness of Lydia's sentiment analysis has examined in several ways:

  - The original *Lydia* sentiment paper [8] identifies significant correlations between our sentiment time series and political poll ratings, sports team performance, and stock indices.
  - We have performed extensive studies demonstrating that *Lydia* sentiment analysis time series can be used to improve the accuracy of movie gross forecasts [18] and also as the foundation for a profitable market-neutral trading strategy for stocks [19].
  - A small study comparing *Lydia* sentiment markup to human coders was performed as in the course of our involvement in the 2008 National Annenberg Election Survey. Although the human coding does not provide sentiment markup which is ideally comparable to Lydia's, the results were still quite favorable.

- *Entity Analysis and Aggregation* – Lydia takes the NLP marked-up documents and processes them in a series of jobs (virtually entirely map-reduce jobs), storing the results in a persistent data structure that we call a *depository*. A Lydia depository includes reference statistics, juxtaposition statistics, article and entity search indices, globally co-referential entity sets, a variety of derived entity classifications, and aggregated statistics for these derived groups.

  The new analysis architecture for Lydia is built on top of the Hadoop [2] implementation of Google's Map-Reduce [7] distributed computation model. As such, it was necessary to construct a dependency management system capable of correctly scheduling the processing of artifacts by map-reduce jobs, especially to correctly manage daily updates of newly processed text. The technical details of this are beyond the scope of this paper, but can be found in [4].

- *Access and Visualization* – Once analysis is complete, a Lydia depository can be accessed through a flexible API which exposes different slices of the data, and provides various visualizations.

3. LYDIA AND NEWS ANALYSIS FOR THE SOCIAL SCIENCES

The analysis provided by Lydia has already proven very useful for research in a variety of areas. In this section, we will describe a number of these applications.

3.1 Lydia in Political Science

Political science is the field that obviously stands to benefit from using our news analysis system, as it is primarily concerned with current events involving entities widely covered in the media. To this end, we have collaborated with political scientists from Stony Brook University and University of Pennsylvania. The general direction of this collaboration is studying the influence of media coverage on electorate opinion, and our system quantifies the media part of this connection.

The Lydia project offers the opportunity to closely examine the relationship between campaign events, public opinion, and media. Many previous studies of media content such as [5,6] have explored an extremely narrow range of news sources. With our new Lydia infrastructure we are now able to analyze roughly 1000 online news sources with an archive spanning four years, starting from November 2004, comprising over 120 million different articles.

One such analysis was performed over 16 months of news, political blog, and TV show transcript sources for the National Annenberg Election Survey (NAES). The NAES is a massive public opinion survey conducted during each of the last three presidential elections by the Annenberg Public Policy Center at the University of Pennsylvania. In each election cycle, the NAES conducts over one-hundred thousand interviews of U.S. adults, examining political issues and attitudes, with an emphasis on the effects of media exposure.

Another study conducted by political science collaborators using the Lydia system is concerned with determining the effect of elite influence on public opinion of foreign policy [9]. That is, there are essentially two models of how public opinion of foreign policy is shaped. The first model is that public opinion is largely based upon relatively tangible, quantifiable measures (events), such as military and civilian casualties, threats to U.S. national security or strategic interest, and the overall prospect of success. The second is that public opinion is largely driven by the influence of a small number of opinion leaders or elites.

Huddie, Johnson, and Lebo [9] examined the effect of media tone, U.S. military casualties, Iraqi civilian casualties, specific important news events, and other effects on the partisan support for

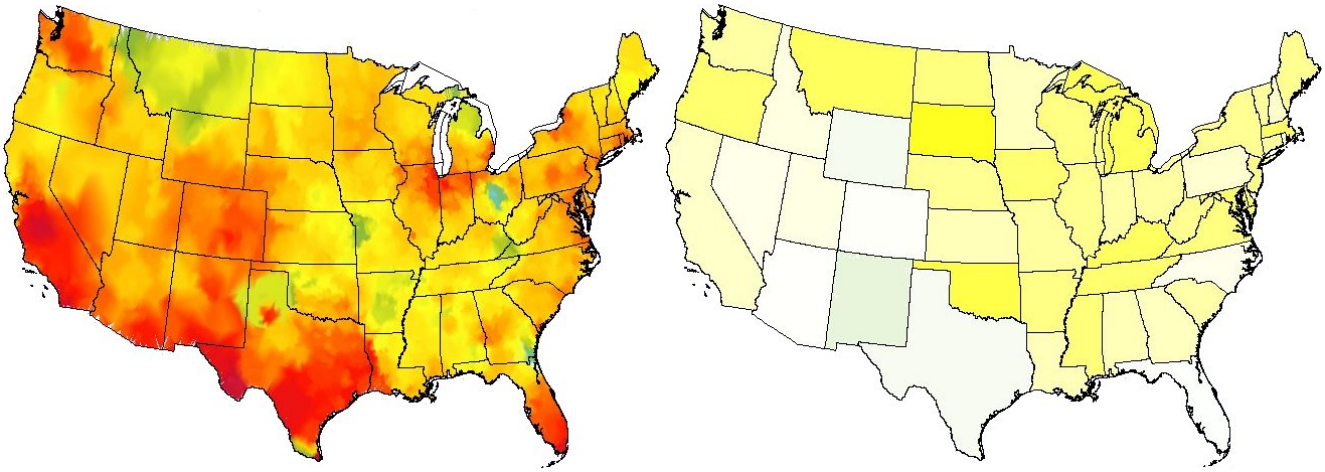

Figure 2: Frequency (left) and sentiment (right) of the Hispanic CEL group in U.S. daily newspapers. For frequency, red reflects the greatest frequency of coverage, while green reflects the least. For sentiment, yellow reflects overall positive sentiment, while white represents neutral sentiment, and green negative sentiment.

the Iraq war over a three year period. Interestingly, results indicate that the strength of these two effects differ across the political spectrum, with republicans being more responsive to specific events indicating the success of the ongoing war, while democrats were comparatively insensitive to these same events.

3.2 Ethnic Bias in the News

Ward, Bautin, and Skiena [17] is a study of news coverage of cultural/ethnic/linguistic (CEL) groups and their interactions using the data obtained from the new Lydia system. It proposes a method for entity nationality detection using juxtaposition data, performs geographic news analysis of cultural groups, examines time series trends in CEL group frequency and sentiment, and quantifies interactions and sentiment between these groups.

Figure 2 shows example figures from this analysis, demonstrating the geographic biases of news coverage across CEL groups. In this example, we examine the differences in volume and sentiment of coverage for individuals in the Hispanic CEL group. We note first that there is a strong regional bias to news volume, which, as we would expect, is highly correlated with the regional variation in Hispanic population density (for a comparison to census data, see [17]). Somewhat more interestingly, we note in the sentiment graph that overall sentiment for Hispanics is strongly inversely correlated with news volume, a trend which does not appear significantly for other CEL groups.

3.3 Lydia in Business and Marketing

Another area in which Lydia's analyses can be put to good use is in the area of marketing research. Presently, we are collaborating with researchers at George Mason University to study a number of questions, one of which is the implications of news sentiment within a marketing context [15]. One simple example demonstrating the relevance of sentiment analysis to this area is shown in Figure 3. Here we see that the sentiment of newspaper coverage surrounding the Hummer brand is strongly tied (correlation 0.6) to gas prices, with a lag time of approximately one year.

Zhang and Skiena [19] studies how company frequency and sentiment data obtained from the new Lydia system reflects the company's stock trading volumes and financial returns. They confirm that the news data is highly informative, as many newspaper coverage variables correlate highly with stock indicators. For example,

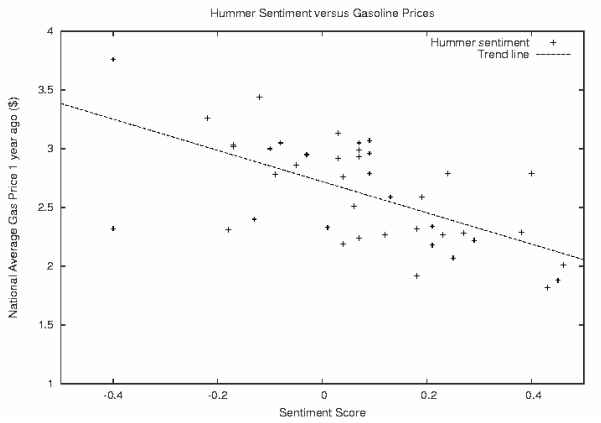

Figure 3: Comparison of Hummer Sentiment and Gas Prices.

reference volume and sentiment subjectivity correlate with trading volume, reference volume correlates with market capital, and most importantly, sentiment polarity correlates with return on investment. Using this data, Zhang and Skiena propose a news-based market-neutral trading strategy which gives consistently favorable low volatility results over the period covered by our news data.

3.4 Lydia in Sociology

Another ongoing collaboration using Lydia data is in the area of sociology [16]; specifically, we are examining questions relating to the acquisition of fame. That is, questions like:

- How is fame distributed?
- Why do some people become famous, and not others?
- What is the ultimate fate of news figures? Does your fate differ by your news area (sports, entertainment, etc.)?
- What role does gender play in the acquisition of fame?

Using the Lydia system, we can now begin to address questions like these, by analyzing the news records of hundreds of thousands of individuals across the country over a period of five years. As

an example, Figure 4 shows the distribution of reference volume of news entities with broadly geographically distributed coverage from the beginning to the end of our U.S. dailies dataset. It is clear that the extent to which an entity is persistent in the news is strongly correlated with its initial level of coverage, and that there is a strong decay effect at work.

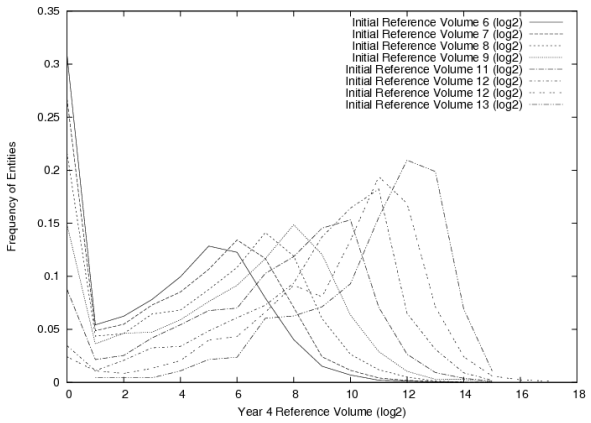

Figure 4: Reference volume distribution of news entities with broadly geographically distributed coverage after four years.

4. ACCESS

Ten processed text corpora (termed depositories in our system) are now available through TextMap Access (`http://www.textmap.com/access`), including:

- *Dailies* – This depository is constructed from a large corpus (over one terabyte) of U.S. and international English-language newspapers, starting November 2004.
- *NAES 2008* – This is the depository of customized analysis for the 2008 *National Annenberg Election Survey*, covering news, blogs, and TV transcripts related to the presidential election.
- *Archival* – This depository is constructed from a longer-term corpus of articles over ten major U.S. daily newspapers, where each source goes back to at least 1995.
- *Historical* – This depository provides analysis from a select set of very long-range news sources providing from 1851 until present times.
- *Pubmed* – This depository of over 17 million Medline/Pubmed journal abstracts permits analysis of trends regarding scientific and medical research, with comprehensive coverage since 1975 and sparser coverage back to 1865.
- *Patents* – This depository of over 3.7 million U.S. patent abstracts charts the scientific and technical landscape since 1971.
- *Supreme Court Decisions* – The depository provides an analysis of all of the almost 60,000 U.S. Supreme Court decisions.
- *LiveJournal* – These blogs were the subject of our original study of blogs [11], and are particularly interesting as a study of randomly-selected blogs.
- *Spinn3r* – Spinn3r [1] is a web-service which provides access to a very large number of blogs and mainstream media sources.

For each of these processed corpora, we provide statistics which can be easily sliced by time and source dimensions. These statistics include frequency, sentiment, and juxtapositions (co-occurrences). Also easily accessible are spatial bias and network visualizations. It is hoped that TextMap Access will make a valuable addition to the social scientist's toolbox.

5. REFERENCES

[1] Spinn3r. `http://spinn3r.com/`.
[2] Apache Software Foundation. The Hadoop Project. `http://lucene.apache.org/hadoop`.
[3] M. Bautin, L. Vijayarenu, and S. Skiena. International Sentiment Analysis for News and Blogs. In *Proc. of the International Conference on Weblogs and Social Media*, Seattle, WA, April 2008.
[4] M. Bautin, C. Ward, and S. Skiena. A scalable architecture for historical news analysis. Submitted, 2009.
[5] J. Box-Steffensmeier, D. Darmofal, and C. Farrell. The endogenous relationship of campaign expenditures, expected vote, and media coverage. In *American Political Science Association annual meeting*, 2005.
[6] H. Brandenburg. Revisiting the "Liberal Media Bias": A Quantitative Study into Candidate Treatment by the Broadcast Media During the 2004 Presidential Election Campaign. In *Proc. of the Annual Meeting of the American Political Science Association*, Philadelphia, Sep 2006.
[7] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proc. of the OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150.
[8] N. Godbole, M. Srinivasaiah, and S. Skiena. Large-Scale Sentiment Analysis for News and Blogs. In *Proc. of the International Conference on Weblogs and Social Media*, Mar. 2007.
[9] L. Huddie, C. Johnston, and M. Lebo. Elite influence, media coverage, and public opinion on the iraq war. In *Midwest Political Science Association 67th Annual National Conference*, 2009.
[10] L. Lloyd. *Lydia: A System for the Large Scale Analysis of Natural Language Text*. PhD thesis, Stony Brook University, 2006.
[11] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *Computational Approaches to Analyzing Weblogs (AAAI-CAAW 2006)*, volume AAAI Press, Technical Report SS-06-03, pages 117–124, 2006.
[12] L. Lloyd, D. Kechagias, and S. Skiena. Lydia: A system for large-scale news analysis. In *SPIRE*, pages 161–166, 2005.
[13] G. A. Miller. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41, 1995.
[14] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers, 2008.
[15] F. Sussan. Personal communicaton, 2009.
[16] C. Ward, S. Skiena, A. van de Rijt, and E. Shor. Sociological news analysis. in preparation, 2009.
[17] C. B. Ward, M. Bautin, and S. Skiena. Identifying differences in news coverage between cultural/ethnic groups. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 3:511–514, 2009.
[18] W. Zhang and S. Skiena. Improving movie gross prediction through news analysis. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, 1:301–304, 2009.
[19] W. Zhang and S. Skiena. Trading strategies to exploit news sentiment. Submitted for publication, 2009.