Visual Analysis of Weblog Content

Michelle L. Gregory, Deborah Payne, David McColgin, Nick Cramer, and Douglas Love
Pacific Northwest National Laboratory
902 Battelle Blvd
Richland, WA 99354
509-375-2824
{michelle; deborah.payne; dave.mccolgin; nick.cramer; douglas.love}@pnl.gov

Abstract

In recent years, one of the advances of the World Wide Web is social media and one of the fastest growing aspects of social media is the blogosphere. Blogs make content creation easy and are highly accessible through web pages and syndication. With their growing influence, a need has arisen to be able to monitor the opinions and insight revealed within their content. In this paper we describe a technical approach for analyzing the content of blog data using a visual analytic tool, IN-SPIRE, developed by Pacific Northwest National Laboratory. We highlight the capabilities of this tool that are particularly useful for information gathering from blog data.

Keywords

Visual analytics, blog analysis, information visualization

1. Introduction

Blogs have become the fastest growing sector of the WWW and have had an increasingly important role in marketing, journalism, and opinion polling [3]. With the growing importance of blogs, there has been an increase in research aimed at techniques for automatic blog analysis and classification. Data mining from blogs presents challenges not typically found in text mining from documents. Blogs can be hosted in a number of different ways, on personal web pages or blog hosting sites, as one long document or individual postings. They contain many more links (to other blogs, news sites, or pages of personal interest) than a typical document, they often contain photos, and blogs can often be repetitions of blogs on other sites [4]. The content of blogs can be very short and informal, containing relatively little content much like email or chat, or none at all in the case of spam blogs. On the other hand, some blogs contain substantial content, much more like a typical news article. These characteristics and variety of blogs pose big challenges when trying to isolate informative content for a specific purpose. As such, the field of blog analytics has grown significantly in the last few years.

Research on blogs ranges from splog identification [9], identifying the source of re-publication of blogs [2], link analysis [11], buzz analysis [12], and social network analysis [1]. There have also been a number of tools developed to aid in visualizing all sorts of blog data. However, these mostly focus on visualizing the connections between blogs or individual bloggers, or on buzz terms or topics over time, see [7] for example. There has been less attention focused on tools for integrated information gathering and analysis of the thematic content of blogs. While market research is often focused on gathering information from blogs about particular companies or products, there is very little in the way of user interfaces designed for exploration and information gathering from large blog datasets.

In this paper we present a methodology for blog analysis using a mature document visualization tool. With this tool, users can harvest blogs (datasets can be static or dynamic, updating with real time information), view them by thematic content, isolate key words of interest, run queries, visualize changes in content over time, or isolate bloggers of interest. Our tool is not designed to capture what is going on in blogs in general in a given time period (capture the "buzz" or track key words over time). Rather, we have designed an analytic environment aimed toward the needs of both business and government intelligence analysts. We anticipate that users of this tool have a task or specific goal and that some information relevant to their task may be contained in blogs. Given the vast amounts of this data type available, we need a tool that can both filter for specific content as well as discover non-explicit relationships.

2. Approach

There are a number of intelligence tools designed for the analysis of large amounts of data by helping to organize documents on topics of interest and placing them in their larger context. These tools have largely been built for single authored, content rich document sets. However, analysts who need a comprehensive understanding of a topic must also have access to new information sources, such as the expanding blogosphere. Rather than providing analysts with a separate tool for distinct data types, we built blog analytics capabilities into an existing application to support analysis across information sources.

2.1 The IN-SPIRE system

IN-SPIRE [6] is a visual analytics tool designed to facilitate rapid understanding of large textual corpora. IN-SPIRE generates a compiled document set from *mathematical signatures* for each document in a set. Document signatures are clustered according to common themes to enable information exploration and visualizations. Information is presented to the user using several *visual metaphors* to expose different facets of the textual data. The central visual metaphor is a Galaxy view of the corpus that allows users to intuitively interact with thousands of documents,

examining them by theme (see Figure 1, below). IN-SPIRE leverages the use of context vectors such as latent semantic indexing for document clustering and projection. Additional analytic tools allow exploration of temporal trends, thematic distribution by source or other metadata, and query relationships and overlaps. By showing topics of interest in context, this tool supports informed exploration and discovery in addition to targeted search.

IN-SPIRE has undergone several recent changes that support visual analysis of blogs, including sentiment analysis and support for streaming content. Some of the visualizations give the user an overview of the data, while others, such as Affect, key in on specific features. These visualizations work in concert with the other tools allowing the user to explore thematic content and isolate data of interest.

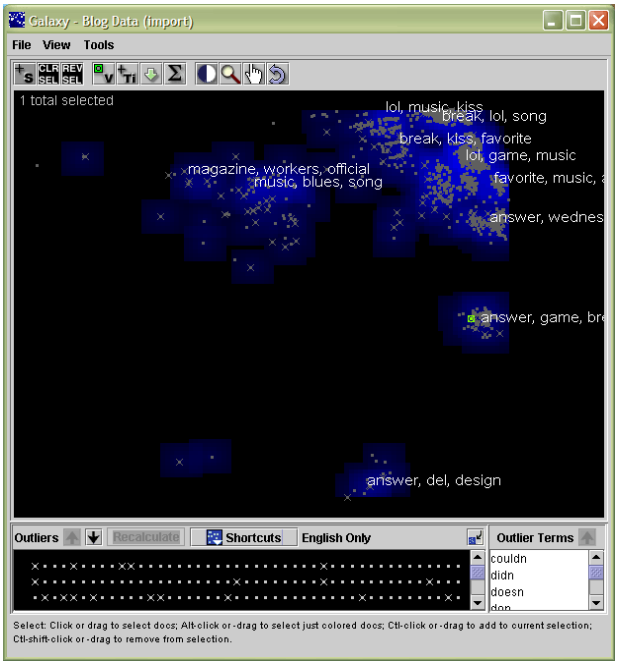

Fig. 1: T*he Galaxy View of a small subset* (*n*=1378) *of the BuzzMetrics data. Dots represent individual blog posts, clustered according to their thematic content. Representative terms are shown next to individual cluster*

Dataset. There are a number of capabilities that IN-SPIRE provides for analyzing blogs. We will demonstrate these capabilities via analyses conducted on the BuzzMetrics dataset made available through http://www.icwsm.org/data.html. We used a random subset of data. We extracted 1000 blog entries from the data on each of the seven days May 1, 2, 4-6, 8, and 9 for a combined total of 7000 documents from the supplied data. A subset of these documents are represented in Figure 1.

2.2 Collection and filtering

IN-SPIRE has a built-in harvester that collects documents or blogs directly from the web, RSS feeds, or from stored static sets. For this paper, we selected a subset of 7000 documents from the sample data set and ran the harvesters on this sample. The harvesters were customized to keep only the blogs text that contained a minimum of 800 characters and text that had a low ratio of the number of characters to the number of links in the document. These settings returned a data set of 2022 documents. While these settings may not be optimal for harvesting all types of blog data, they did provide a first pass at isolating blogs with a richer text content (for example, excluding those blogs that only contained links or pictures). The text-poor blogs that were culled offered little fodder for the deep text analysis users perform with IN-SPIRE. These settings did not exclude any blogs based on language, thus our final blog set contained blogs from multiple languages.

2.3 Multi-lingual data

One of the lesser explored issues of blogs analytics is finding ways to extract content from multiple languages at once. Our suite of tools designed for multi-lingual datasets includes integration with third party **language detection** and a number of third party machine translations systems in order to **translate cluster labels**, **perform multi-lingual queries**, and **translate blogs of interest in full**. We have integrated third party translators for over 40 languages and third party software for language identification. Datasets compiled with language detection allow IN-SPIRE to automatically select the most appropriate translator for each document.

To explore a foreign language dataset, the system clusters the documents in their native language (with no pre-translation required). A user can then view the cluster labels, or peak terms, in the native language, or a translated version, using locally available machine translation software. For this study, we used Systran [10]. The user can then explore the clusters to get a general sense of the thematic coverage of the dataset. Isolating clusters relevant to their interests allows them to re-cluster to show more subtle themes differentiating the remaining documents. If they search for particular words, the clusters and translated labels help to distinguish the various contexts in which those words appear. Finding a cluster or document of interest, a particular document or set of documents can be viewed and translated on demand. This avoids the need to translate the entire document set, so that only the documents of interest are translated. The native text is displayed alongside the translation at all stages.

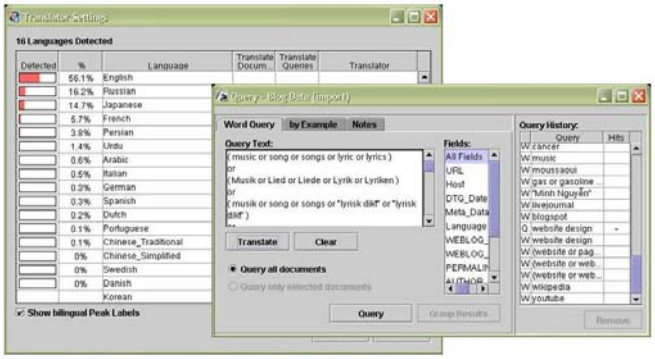

Fig. 2: *Language detection and multi-lingual query tools*

The **translator settings tool** allows one to view percentage of each language, as well as control the active translator. Figure 2

illustrates that our document collection contains 56% English blogs, 16% Russian, 15% Japanese, and so on. The **multi-lingual query tool** allows one to perform a query in their native language (English, in this case), and translate it to any of the other languages in the dataset. The English query is translated into all selected languages and the query results are interleaved. As such, results of the query will include documents from all languages in which the query found matches. For example, in Figure 2, the query was a Boolean OR query with terms related to music (e.g. music, song, lyrics). The query resulted in a group of blogs that all had something to do with songs or music, independent of the original language of the blog. Figure 3 shows query groups in the Correlation Tool in order to investigate the relationship between documents from blog language (Russian, Farsi, and French) and cross-language queries on television, music, and movies. It is evident from the middle column that music is a consistently common topic across languages.

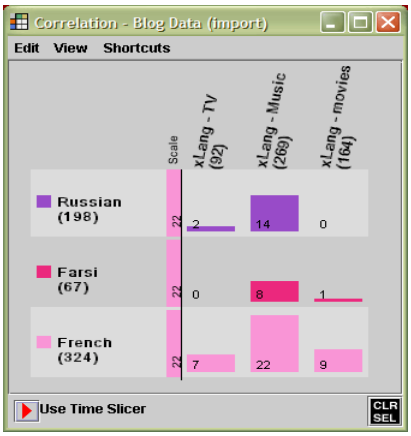

Fig. 3: *Correlation of blog language by queries on media type*

This system was originally designed for monolingual datasets but the document representation is language-independent. Thus, our sample mixed-language dataset clusters each document based on its native language and the Galaxy often naturally "clumps" along language boundaries, helping the user to interpret query results and triage the data.

2.4 Query tools

One of the central tools of the IN-SPIRE application is the ability to form complex queries. For example, posts within a data set can be queried to reveal content within the blog or within any of the fields of the data such as title, author, or source URL. Our query tools support both **Boolean** search and **Query By Example** (QBE). In addition, our tool can link to **other question answering systems** to support natural language questions and semantic query expansion based on the data set [8].

Queries can be saved as groups of result documents. IN-SPIRE groups are the basis of several other powerful analysis tools such as Correlation (Figure 3), Affect measurement, hypothesis tracking, and more. In Figure 4, a metadata query was used to isolate posts from an individual blogger. The resulting group is shown over time relative to the whole data set, revealing patterns of activity. This blogger posted on 4 non-contiguous days, with several more posts on May 2, in fact, 22 of his 39 posts are from that day. Scanning the titles and text from the day reveals that it is near the first day of an academic term.

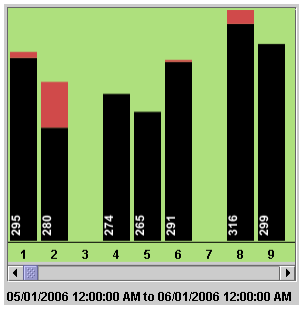

Fig. 4: *Time Slicer tool showing the total number of documents per day* (black) *and the queried blogger's posts* (red)

Queries can also be stored in the Triage tool that allows users to store not only the query terms of interest but also a set of operations to filter the combined results to the most useful documents. For example, Figure 5 shows a set of queries, or *nodes*, on the left that represent groups based on host site (such as Live Journal or Blog Spot), cross-language queries for television- and gasoline-related terms. The node on the right combines the blog posts from the connected nodes on the left, representing posts from Blog Spot about gasoline that have been added to the dataset recently. Selecting a Triage group selects matching documents and portrays the overlap with all other Triage groups. Together all these nodes represent a Triage *network* that can be reapplied to other data. In streaming datasets, the network updates every time new documents are added. As the dataset is updated the underlying tools such as the Triage tool provide the user with up to date analysis of the changing blog content.

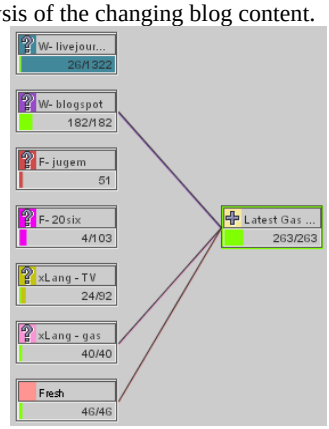

Fig. 5: *Triage tool showing queries on the left and a Combo node on the right*

2.5 Affect analysis

The affective content of blog posts can be a powerful tool in discovering not what people are talking about, but also the sentiments that frame their statements. IN-SPIRE supports affect measurement at a document level (in this case blog posts) that can then be analyzed using groups. Affect is measured using a lexicon approach for 5 different affect pairs: *positive/negative, virtue/vice, pleasure/pain, power coop/conflict,* and a smaller subset of *positive/negative*. In general, these axes are loosely based on the General Inquirer lexicon. Each axis is represented by a different color (pairs are different shades of the same color) on a rose plot,

shown in Figure 6 (for more on a description of affect measurement and visualizations, see [5]). Each rose plot represents the distribution of affect terms for a particular group.

To compare how affect differs amongst different groups (whether those groups be based on blog sites, topics, or individual bloggers) one just needs to compare the size of the petals for the different affect axes across groups. For example, at the time of BuzzMetrics data collection, Zacharias Moussaoui was being sentenced for his minor role in the September 11th terrorist attacks. Several bloggers wrote about the sentencing. The top left-most box in Figure 6 represents the affect distribution in these blogs compared to the music-related blogs we saw earlier, represented in the bottom left-most box. The top two "petals", in bright red and dark red, represent positive and negative, respectively. It is clear from the plots that the Moussaoui blogs contain many more negative terms than the music blogs.

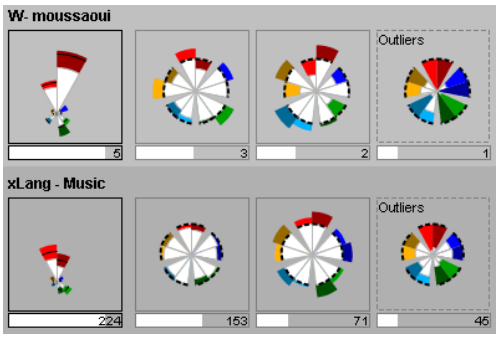

Fig. 6: *Affect Distribution of two blog topics*

3. Summary

In this paper we have introduced an application for comprehensive blog analytics. The application is designed for data gathering and information discovery, and has recently been applied to blogs. To date we have not performed any formal evaluations of the tool with this data type. The diverse nature of blogs makes it difficult to identify the most useful metrics. We expect the strongest evaluation may come from studying users performing realistic analysis tasks and we hope to accomplish a pilot user study in the near future.

Using the tool with blogs has suggested a number of targeted improvements. Foremost, we are interested in developing representations of the often sparse content of blog data that allow both blog data sources and traditional document types to be represented in the same collections. Currently, the number of unique terms in more traditional text collections masks the contribution of the sparser blog data types. Also, while we believe multilingual capability is one of the most promising aspects of our approach, there is room for improvement. Our current implementation allows one to isolate blogs by language and to see the sub-themes within language groups, but it would also be helpful to be able to have all clusters reflect thematic content in the user's native language, even if it requires multiple translators simultaneously. In addition, the machine translation software we use—which relies on language models built from content rich data sources—needs much improvement to adequately translate content characteristics of blogs, including slang terms, misspellings, etc. Lastly, in our approach the blog data is gathered from specific RSS feeds, data web pages, or static data collections. This methodology inherently includes some sort of filtering. It seems worthwhile to include other filtering approaches at the front end, such as integrating our harvester with a "buzz" approach or including a splog identifier, to ensure targeted, cleaner source files.

References

[1] Adamic, L.A., and Adar, E. Friends and neighbors on the Web. *Social Networks*, 2003. 25(3):211-230.

[2] Adar, E., and Adamic, L.A. Tracking Information Epidemics in Blogspace. In *Proceedings of Web Intelligence*. (Compiegne, France, Sept. 19-22, 2005).

[3] Blood, R. Weblogs: A History and Perspective. Rebecca's Pocket. 07 September 2000. 25 October 2006. http://www.rebeccablood.net/essays/weblog_history.html.

[4] Cohen, E., and Krishnamurthy, B. A Short Walk in Blogistan. *Computer Networks*. 2005. 50(5):615-630.

[5] Gregory, M., Chinchor, N., Whitney, P., Hetzler, E., and Turner, A. User-directed Sentiment Analysis: Visualizing the Affective Content of Documents. *ACL workshop on Sentiment Analysis Workshop*. Sidney, Australia.

[6] Hetzler, E., and Turner, A. Analysis Experiences Using Information Visualization. *IEEE Computer Graphics and Applications*, 2004. 24(5):22-26.

[7] Lento, T., Welser, H.T., Gu, L., and Smith, M. The Ties that Blog: Examining the Relationship Between Social Ties and Continued Participation in the Wallop Weblogging System, In *Proceedings of the third annual workshop on the weblogging Ecosystem: Aggregation, Analysis, and Dynamics*. (Edinburgh, Scottland, May, 2006).

[8] McColgin, D., Gregory, M., Hetzler, E., and Turner, A. From Question Answering to Visual Exploration. In *Proceedings of the ACM SIGIR workshop on Evaluating Exploratory Search Systems.* (Seattle, USA, August 10th, 2006). pp. 47-50.

[9] Salvetti, F., and Nicolov, N. Weblog Classification for Fast Splog Filtering: A URL Language Model Segmentation Approach. In *Proceedings of HLT-NAACL 2006: Human Language Technology Conference*. (New York City, NY, USA, 2006).

[10] Snellart, J., Yang, J., and Rebollo, A. SYSTRAN Intuitive Coding Technology. In Proceedings of the MT Summit IX (New Orleans, Louisiana, USA, 2003).

[11] Thelwall, M. *Link Analysis: An Information Science Approach*. Academic Press, 2004

[12] Yi, J. Detecting buzz from time-sequenced document streams. In *Proceedings of the IEEE conference on e-Technology, e-Commerce and e-Service, 2005. (EEE '05)*. (March 29 - April 1). pp. 347- 352.