Visual Analysis of Weblog Content

David McColgin, Nick Cramer, Michelle L. Gregory, Deborah Payne, and Douglas Love
Pacific Northwest National Laboratory
902 Battelle Blvd
Richland, WA 99354
509-375-2128
{dave.mccolgin; nick.cramer; michelle; deborah.payne; douglas.love}@pnl.gov

Abstract

In this demo we present an analytic tool that provides a comprehensive approach to weblog analysis. We have combined ingest tools that have been designed to work with the special characteristics of blog data with a mature visual analytic tool designed for in-depth document analysis. Using a subset of the Buzz Metrics data, we demonstrate how this combination of tools provides users with the capability to analyze blogs in multiple languages, investigate changes over time, and investigate the affect of blogs.

1. Introduction

There are a number of 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, users need a comprehensive understanding of a topic, thus 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.

The IN-SPIRE system, described in more detail in [1], is a visual analytics tool designed to facilitate rapid understanding of large textual corpora. We use this tool as a basis for blog analytics.

2. Blog analysis in IN-SPIRE

In this section we describe how we use this tool to aid in providing a comprehensive analysis of blog data. We used the Buzz Metrics data provided through the conference web page.

2.1 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 demo, we created keyword subsets of documents from the sample data set and ran the harvesters on the result. For example, we ingested all of the blog posts that mentioned "HD-DVD", among other keywords. For the majority of the demo, we used blogs that matched the search term United 93, beginning with 3336 documents for analysis.

2.2 Galaxy view and outliers

One of the central visualizations in IN-SPIRE is the Galaxy View in which a user can visualize all the documents in a single space clustered by content. Documents in this view are clustered by topics, determined by the mathematical signatures of each document.

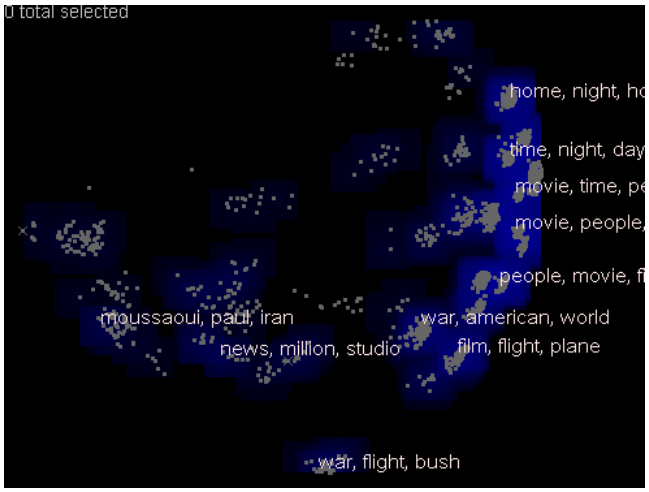

Fig. 1: *The Galaxy View of the refined United 93 subset* (*n*=3155) *of the BuzzMetrics data. Dots represent individual blog posts, clustered according to their thematic content. Representative terms are shown next to individual clusters*

The movie United 93 was released the week that the Buzz Metrics corpus was collected. The Galaxy makes it easy to see the main topic groups of the documents and allowed us to quickly remove uninteresting clusters, such as documents that were predominantly about box office totals. We can see in Figure 1 that the resulting subset includes document clusters pertaining to the trial of Zacarias Moussaoui – a 9/11 conspirator that was convicted during the same week. A cluster at the bottom contains topics related to the Bush administration and the war on terror.

2.3 Query tools

One of the central tools of the IN-SPIRE application is the ability to form complex queries (see [1] for more information). Once might query the document set for the source of the blog content, or perhaps the individual bloggers. Queries can be saved as groups of result documents. IN-SPIRE groups are the basis of several other powerful analysis tools such as Correlation in Figure 2, which shows keyword hits by host. It is evident that Moussaoui is talked about more by Blog Spot users than Live Journal or Xanga, even though it has only a third of Live Journal's authors. No one on Xanga mentioned Al-Qaeda.

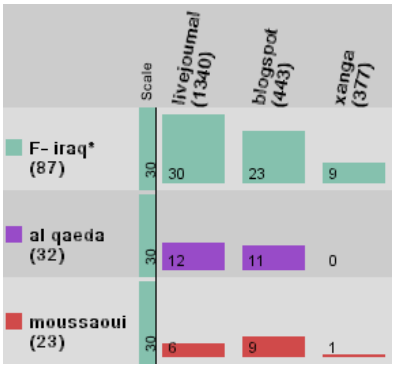

Fig. 2: *Correlation of blog posts by blog site*

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 3 shows a set of queries, or *nodes*, on the left that represent groups for the blog hosting site (such as Live Journal or Blog Spot), keyword queries on people, other movie titles, and more. The nodes on the right combine the blog posts from the connected nodes on the left. Thus, the rightmost node contains posts from any of the hosts shown that also discuss Bush and Iraq. 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 is stored independently of the data and can be reapplied to any other dataset collected. In streaming datasets, optionally fed by RSS updates of new posts, 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. This is especially useful when monitoring activity on more than 1 blog site—the triage network means user strategies for finding key information are automatically reused.

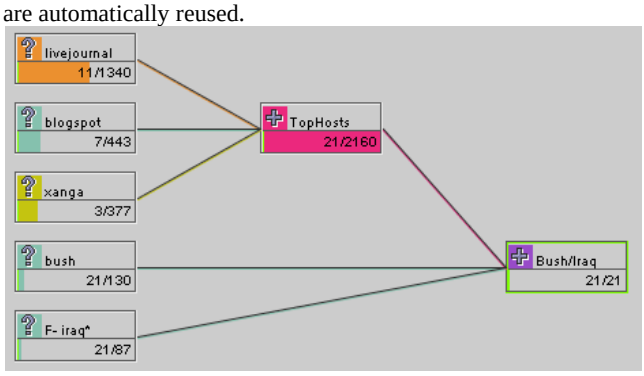

Fig. 3: *Triage with queries at left and a combo at right*

2.4 Affect analysis

We can use the affect analysis of blogs to find out if people liked the movie and the affect visualization of all posts shows that the reactions are overall positive.

However, we can also do a bit of social analysis. For example, we might want to see if different blog sites have differing opinions on the movie, perhaps because a different demographic use different sites. Figure 4 represents a comparison of groups of posts in which we can see that while all sites have more positive than negative affect, posts from My Space are more profusely positive.

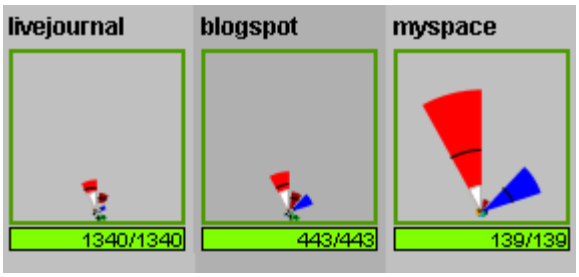

Fig. 4. *Affect distribution of posts broken down by three hosts* (The prominent red wedge at the top represents positive affect)

2.5 Time slicer

The groups in IN-SPIRE also work in concert with many of the other built in capabilities, such as the Time Slicer (Figure 5). In Figure 5, we can see how the number of posts about this movie changed over the course of these 12 days (as indicated by the total length of the bars). Monday the 1st has the most posts as it is after the premiere weekend. May 7th is also a spike as weekend moviegoers respond to the movie online. The orange portion of the bar represents a group of posts identified from Affect as more positive than expected, and shows a fairly even distribution.

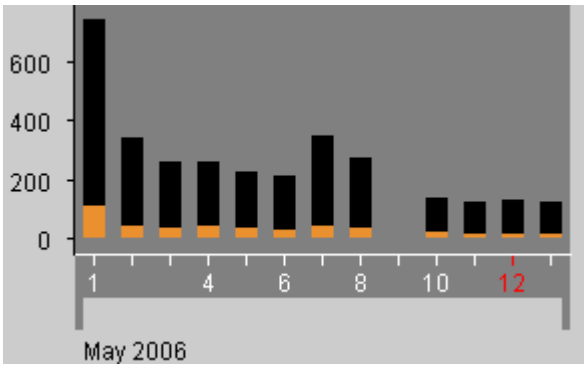

Fig. 5: *Time Slicer tool showing the total number of documents per day* (black) *and the documents with highly positive affect* (orange)

3. Summary

In this paper we have introduced a tool for comprehensive blog analytics. Using this tool, we have examined the 3336 blogs on United 93 the week it came out in 2005. The Galaxy shows interesting facets of the information we can explore. Query and Triage allow for targeted investigation and automatic reuse of searches. With Time Slicer, we can see that the reduction in posts over time and the increase on the weekend. Affect shows reactions to the movie and patterns over any other groupings we have made. The most insight comes from using the tools in concert towards a larger analytic goal. The tools described here are complemented with support for foreign language analysis, "steering" visualizations towards topics of interest, and more.

References

[1] Gregory, M., Payne, D., McColgin, D., Cramer, N., Love, D. Visual Analysis of Weblog Content. ICWSM '07. Boulder, CO