Weblogs as Market Indicators:

Tracking Reactions to Issues and Events

Richard Tong

Tarragon Consulting Corporation
1563 Solano Avenue, #350
Berkeley, CA 94707
rtong@tgncorp.com

Mark Snuffin

DataNaut LLC
4400 MacArthur Blvd., Suite 102a
Washington, DC 2007
snuffin@datanaut.com

Abstract

We have an ongoing interest in the large-scale analysis of consumer-generated media, of which weblogs are currently of special concern. In this position paper, we describe how our data gathering and analysis system, called T2™, is being used to collect weblog entries and analyze them for reactions to issues and events in the commercial and government arenas. The paper includes preliminary lessons learned in trying to apply our sentiment analysis techniques to the highly informal language usage found in weblogs.

Operational Motivation

The blogging phenomenon is now well documented, with Technorati, for example, reporting that "the blogosphere continues to double every 5.5 months," and that they are "tracking about 900,000 blog posts created every day."

This outpouring of commentary on anything and everything of interest to individual consumers adds yet another dimension to the set of large scale "conversations" that the Internet supports.

The challenge is to see whether this on-line material tells us anything of value about the reactions that consumers have to issues and events. In particular, whether weblogs provide us with market indicators that can enhance and complement more traditional marketing research techniques.

As in our previous work (Tong and Yager, 2004), we are primarily concerned with large-scale behaviors rather than analysis of any specific posting, and with the general sentiment expressed rather than the raw mentions of topics. This kind of analysis is of potential value in both commercial and government sectors. It can be used to gauge reaction to new products, or assess the impact of an advertising campaign, as well as to track attitudes towards US foreign policy, or understand the sentiment "on the street."

Collecting and Processing Weblogs

From one perspective, weblogs for us are just another online source. We collect them either directly using RSS/Atom feeds, or by using one or more of the blog search engines.

From another, though, the highly idiosyncratic nature of weblogs, both with respect to language usage and posting structure, makes the application of standard text analysis techniques problematic, and has led us to explore the use of simpler pattern matching techniques.

The goal is to balance the competing demands of accuracy and processing speed, while maintaining our ability to say something meaningful about the sentiment being expressed as we aggregate across time and postings.
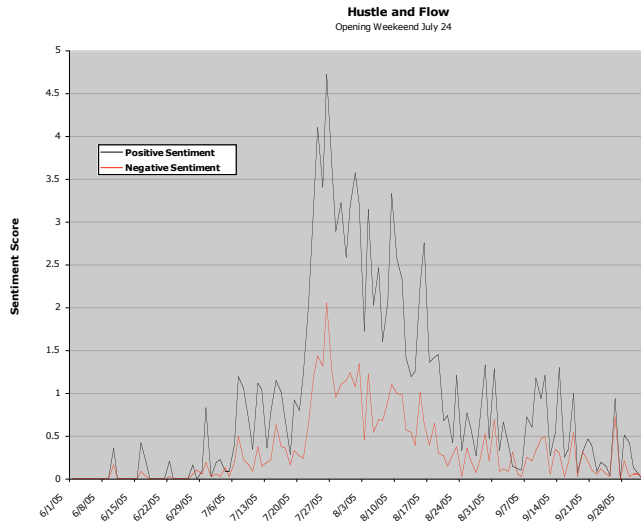
In the weblog experiments reported here, we have focused on movies, since they are a common theme in weblog postings and engender plenty of evaluative language. In addition, movie releases are well-defined events with public-domain information, such as casting, advertising budgets and box-office receipts, all of which can be used to interpret and validate the signals we extract from the weblogs.

The basic method we use to analyze weblog postings first applies a series of segmentation rules to the original webpage. The objective is to identify content bearing elements so that subsequent processing is applied only to those page elements that are the original contribution of the weblog author.

After conversion to XML, the segmented page is analyzed using a sequence of patterns that look for affective language use in combination with various movie descriptors. Each pattern is differentially weighted, and the evidence from each is combined to give an overall sentiment score.

Our experiments to date have focused on repurposing families of lexico-syntactic patterns that we developed for other online forums, such as bulletin boards. The key technical challenge is to define an appropriate context for the application of the patterns, and, so far at least, we have been primarily exploring the use of variable length windowing techniques.

The figure below shows example sentiment timelines for the movie *Hustle and Flow*, which was released on July 22, 2005. The upper curve (colored black in the original) shows the positive sentiment. The lower curve (colored red in the original) shows the negative sentiment.

Hustle and Flow
Opening Weekeend July 24

We see that these timelines follow a familiar pattern; increasing chatter before the opening, a spike at the opening weekend itself, then a decreasing trend as fewer and fewer theaters show the movie.

The sentiment scale is based on the idea that both positive and negative sentiment range over the unit interval, and that a given posting thus contributes at most 1.0 to the sentiment scores. On a daily basis, the overall sentiment score is just the sum of the sentiment scores for postings on that day.

Weblog Analysis

In this current project, we looked at weblog postings on a selection of movies released in 2005, and restricted our analysis to those in English.

A review of the corpus we collected suggests that, for this domain at least, we can divide postings into two main groups. Those that are substantive movie reviews, and those that mention the movie "in passing."

An example of a posting in the first group is from blogspot.com on July 30, 2005 in which the blogger is writing a review of the move *Hustle and Flow*. The posting contains usage like "the best movie that I have seen this year," "the soundtrack is tight as hell," and "I recommend this movie highly." Such postings look just like material from other sources we have worked with, and are thus amenable to the application of similar analytic techniques.

An example of a posting in the second group is from blogsome.com on August 9, 2005 that starts:

> No seriously, holler at ya girls…"whoop that trick" (a la Hustle and Flow) if u have to b/c I give up.

where the movie is used a cultural reference and in a highly informal linguistic context.

Another example from the second group is more typical. In a posting from livejounal.com on August 29, 2005 we see:

> In Chicago, my mom dragged me to Hustle and Flow. In Boston, it was 40 Year Old Virgin. Entertaining the second time, but not as much so at all. Seth Rogan is hottttt.

where the reference simply marks one activity among a series of activities that this posting describes, but is not really the point of the posting.

Our approach to analysis is first to locate postings that contain a mention of the target (in this example the movie *Hustle and Flow*). For each of these postings, we generate a series of text windows of different widths centered on the target mention. Then in these windows we look for both domain specific and evaluative language usage. That is, we look for language that is specific to movies, such as the names of actors and directors, and various aspects of movie making, such as the soundtrack and the editing. We also look for affective language, both with reference to personal state, as well as to movie characteristics.

We have been experimenting with windows of size 20, 100, and 250 words, and using lexically-based patterns that we originally developed for other online movie forums. The strategy is to make multiple passes to locate collocations of patterns, and then differentially combine the evidence we find into an overall sentiment score.

Given the often irregular use of punctuation and markup in weblog postings, we make no attempt to identify anything that looks like a sentence or paragraph boundary, and rely entirely on these proximity surrogates to define regions of interest.
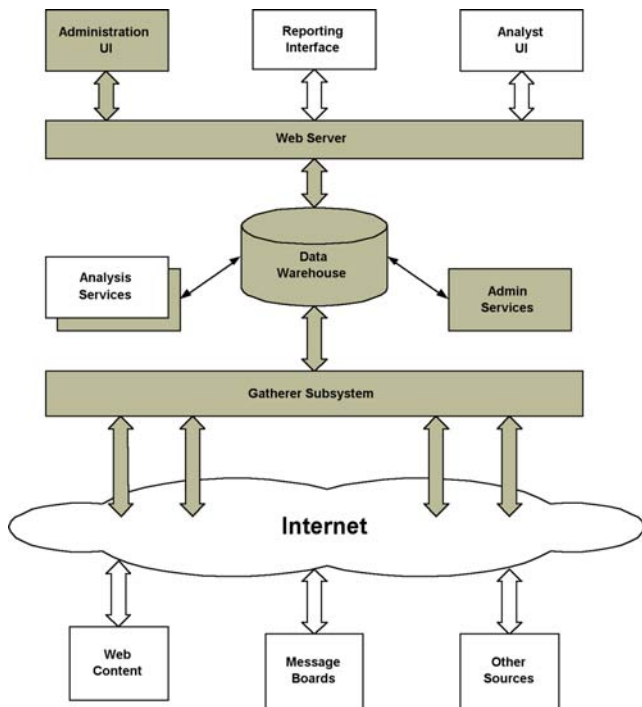
To date, we have not experimented on the weblog data with any NLP techniques, such as shallow parsers. Our experience with other informal online sources is that the value we get from the use of more sophisticates language analysis techniques depends critically on the insights we are trying to gain from the data.

Since our current focus is on large-scale behaviors and the market signals they can give us, we are working on the assumption that the kinds of trend analysis we are developing are relatively insensitive to point failures of language interpretation. Future experiments will attempt to test this conjecture.

T2™ Infrastructure

The T2™ infrastructure is a scalable, configurable content acquisition system that is designed to collect and manage large amounts of open source material, such as weblogs, as well as that found on websites, message boards, mailing

lists, and other Internet forums. The figure below gives an overview of the architecture.

T2™ is a totally automated approach to data gathering and analysis that combines smart web crawlers, data warehousing technology, fully configurable data processing and analysis workflows, and interactive report generation.

In the base configuration (shown shaded in the figure) the T2™ infrastructure includes a set of standard "Analysis Services," including a generic query language and search capability, families of meta-data extractors and taggers, content analyzers that use fuzzy regular expression techniques, and document de-duplication algorithms.

Specialized modules, such as advanced sentiment tracking algorithms, and time-series analysis tools can be inserted into the processing architecture as needed to meet specific application requirements.

Commentary

Blogging is a phenomenon that certainly deserves our attention. On one level it represents a perfect expression of the power of the Internet to democratize the publication and dissemination of commentary and opinions. On another, it provides us with a massive amount of naturally occurring text that presents novel challenges to our existing arsenal of language processing tools and technologies.

From a business perspective, though, we need to ask what it is about weblogs that is of value to us. The fact that people are posting in an easily accessible public forum obviously presents us with a unique opportunity. At the same time, however, we need to focus on what "signal" we want to extract from the background "noise."

A movie studio trying to assess the impact of pre-release advertising, has a different goal from a consumer product company trying to get insight into the market's reaction to a new product feature. In turn, these goals are distinct from that of a political organization that is trying to build support for a policy issue.

In all these scenarios, what we currently lack is any notion of a benchmark against which to calibrate the responses we see. What we want to know is whether the extracted signal correlates with the number of tickets sold, or gives us insight into potential customer support issues, or can predict the outcome of an election.

A critical challenge for researchers interested in the computational analysis of weblogs, and, by extension, other consumer-generated media, is to understand the motivations behind personal expression on the Internet, and how this relates to actions in the real world. Processing the content of postings is certainly part of the puzzle, but we need to do the interpretation in the appropriate social and cultural context if we are to create tangible value.

References

Technorati Weblog. State of the Blogosphere, August 2005, Part 1: Blog Growth. Posted by Dave Sifry on August 02, 2005.
http://www.technorati.com/weblog/2005/08/34.html

Technorati Weblog. State of the Blogosphere, August 2005, Part 2: Posting Volume. Posted by Dave Sifry on August 02, 2005.
http://www.technorati.com/weblog/2005/08/35.html

Tong, R. M., and Yager, R. R. 2004. Characterizing Attitudinal Behaviors in On-Line Open Sources. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, AAAI Technical Report SS-04-07, Qu, Y.; Shanahan, J.; and Wiebe, J. eds.