Temporal Text Mining

Matthew Hurst
Intelliseek Inc./BlogPulse.com
mhurst@intelliseek.com

Abstract

Text mining often involves the extraction of keywords with respect to some measure of importance. Weblog data is textual content with a clear and significant temporal aspect. Concepts in this data, therefore, can be described temporally in terms of the patterns present in their time series. This paper describes a system which uses models of time series to rank keywords from a corpus of weblog posts.

Introduction

The growth of online personal media (weblogs, message boards, usenet, etc.) has resulted in the emergence of a new type of business and marketing intelligence solutions space. Online data is mined to determine a number of issues important to many corporate functions, including brand monitoring, alerting, competition tracking, sentiment mining, customer care and so on.

These systems tend to have either a prescriptive structure: a number of topics are defined and the data is mined to provide a corpus, or an atemporal one: a corpus of documents is mined for keywords, phrases, entities and relationships. In the former case, we may observe the temporal pattern of a known term or class of documents. In the later, we are capable of extracting discriminating features (such as keywords).

At best, alerting mechanisms use simple temoporal models that include either a sudden change in some rank, or the crossing of a threshold in order to determine the importance of a term over time.

However, for many applications it is important to discover which terms (or which concepts to provide a more general framework) are trending in a given way. It a process for discovering concepts whose temporal profile fits a certain pattern, not a process for discovering the trends of known concepts.

This paper describes a system which is capable of mining blog data for terms which fit certain temporal models.

Time Series

A time series is a complete ordered sequence of periods each of which has a value. For a given series T, we use $t(i)$ to denote the value of time period i. We use T_{bg} to indicate a *background* time series. This time series represents the entire corpus of documents, all others being subsets of this corpus.

The value, $t(i)$ may be an absolute count, or - as in this paper - a normalized value with respect to T_{bg}. Normalizing T against T_{bg} simply produces a time series for which $t(i)$ is equal to $t(i)/t_{bg}(i)$.

In the implemented system, a time series is derived from a sorted set of time data representing time stamped events. This allows for the derivation of time series of arbitrary granularity (hour, day, week, month) via a trivial linear scan.

Temporal Models

One way to describe time series is to describe a number of simple elements and then summarise a time series as being composed of a sequence of these elements: a model. One can imagine a simple (regular) grammar expressing models such as 'linear trend up, followed by a plateau and then an exponential drop off.'

When fitting such patterns to a time series, one has to consider the *segmentation* and some *measure* of the fit of each element - we have an optimization problem suitable for a dynamic programming solution.

In this initial exploration, we take a somewhat simpler approach. We define two simple elements:

linear a straight line.

burst an initial flat component followed by an acute jump in the last period.

Each element is captured by a procedure which fits the model directly for a given time series, producing a measurement of the fit, as well as in the case of the linear element, a qualification indicating the direction of the trend (up or down).

The procedures are:

Regression linear regression is used to return:

m : the gradiant

r^2 : the correlation coefficient.

m allows us to classify increasing and decreasing trends.

Burst The burst score, $b(T, p, q)$, is computed as follows:

$$\frac{t(q)}{(\sum_{i=p}^{q} t(i))/(1+q-p)}$$

This captures the notion of a sudden jump in values, the ideal being a flat line with an infinite value in the final time period.

Each of these procedures allow us to derive a number of metrics. Firstly, we define the *data density*, $d(T)$, of T as:

$\sum_{0}^{n-1} 1$ if $t(i) > 0$; else $0/|T_{bg}|$

the number of time periods with non zero values divided by the length of the background time series.

The metrics are:

linear up : an upwards trend fitting a straight line, m must be greater than 0, the metric is $d(T) * r^2$.

linear down : a downwards trend fitting a straight line, m must be less than 0, the metric is $d(T) * r^2$.

final burst : a relatively stable trend with a burst in the final period, the metric is $b(T, 0, n-1)$.

maximum burst : $\max(b(T', 0, p)) : p < n$

In what follows, we describe how metrics are delivered for a corpus of weblog posts and then illustrate the value delivered by these metrics by presenting a case study over a recent corpus of weblog data.

Methodology

The algorithm proceeds in the following manner:

1. The corpus of documents is scanned linearly.
2. Date counts are maintained for each token at a per document level.
3. A time series representation is derived for each token.
4. Each time series is bucketed to a specified granularity (day, week, month).
5. The background time series (T_{bg}) representing the message counts per day is created and similarly quantized.
6. Each token time series is normalized against T_{bg} to provide a normalized time series.
7. Each timeseries is scored against each metric.

Once all the time series have been scored, the system presents the user with a table of results allowing them to rank the tokens by any of the metrics. The user may then inspect the results utilizing a time series visualization.

Case Study

Evaluating trend mining technology is something of a complex problem. One could explore the impact they have on a certain analytical process - for example did they alert the uesr to some trend that they were not aware of and which had a large significance? Alternatively, they could be evaluated in the context of some other task - for example, did they improve the ranking of queries in a search engine (DJ04)?

Here, we simply present a case study to illustrate the typical behaviour of the mining system.

Data Collection

18, 417 blog posts were collected from the BlogPulse index, proportionally sampling between July 1st 2005 and, September 19th 2005 using the query term 'bush.' This data set has a clear temporal event in it caused by the hurricane Katrina crisis and the reaction of the blogosphere both to this disaster and the way in which the Bush administration handled it.

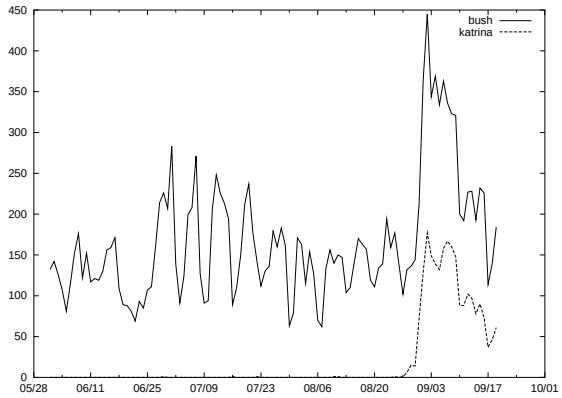

Figure 1: A sample of blog posts for the query term 'bush' plotted over time.

Figure 1 shows the time series for this data set. There is a clear spike related to Katrina peaking on the 2nd of Septmber (Katrina hit the coast on the 29th of August).

Focusing on the 12, 283 document found in the period between June 5th and August 28th which represents a period during which there was relatively level volume of posts containing the term 'bush', we can begin to explore the results of the mining system described above. This period has been selected as it represents an archetypal trend mining problem - there is no clear event or trend in the corpus of documents other than a clear periodic posting trend common to all samples from the blogosphere caused by the fact that bloggers tend to publish content during the week, not on weekends.

Table 1 shows the top 20 terms mined using the linear up metric. It is not surprising to see the two terms 'august' and 'aug' appear in this list - the time period ends during the month of August. Nearly all the other terms, however, provide insight into the issues that were trending up during that time period.

secular part of the debate on Iraq includes a number of issues involving secularism, from the secular nature of Hussein's regime to the goal of having a strong secular Iraq as a foothold for democracy in the middle east.

tour refers to the Tour de France (between X and Y), Cindy Sheehan's bus tour and mentions of the Tour de Crawford - a satirical reference to Bush's practice of biking around his country estate.

sharon refers to Israeli Ariel Sharon and the evacuation of Gaza.

gasoline refers to rising gasoline and oil prices. It is interesting to note that in this time period, none of the posts

Token	Linear Up
secular	0.92
lived	0.86
august	0.85
regional	0.84
mistake	0.83
southern	0.83
gasoline	0.82
62	0.82
protesters	0.81
condition	0.81
marine	0.81
mountain	0.81
tour	0.80
hurt	0.79
sharon	0.78
highway	0.77
casey	0.77
vacation	0.77
israelis	0.77
landed	0.76

Token	linear down
powerful	0.88
street	0.88
social	0.85
direction	0.84
june	0.84
prisoners	0.83
promises	0.82
patriot	0.82
chemical	0.81
tony	0.81
prime	0.80
witnesses	0.80
downing	0.80
justification	0.79
minister	0.79
gitmo	0.78
torture	0.78
define	0.77
jack	0.77
seeing	0.77

Table 1: Top 20 results for the metrics linear up and linear down.

mentioning the term 'gasoline' mention the term 'katrina.' In the period after the hurricane hit, 86 of the messages mentioning 'gasoline' also mentioned 'katrina', and of those, 44 also mentioned either 'price' or 'prices.'

vacation refers to Bush's vacation, particularly in contrast with the general problems his administration faces and the growing anti-war sympathy focused on Cindy Sheehan and Camp Casey.

In summary, by extracting terms that are trending upwards, we can discover that in this relatively flat period of postings about Bush, a number of issues are coming to the fore including anti-war sentiment. the accountability of the Bush administration as well as a significant change in the perennial middle east situation.

Figure 2 plots a selection of the upward trending terms. Note that these plots are normalized and the values are percentages of the background corpus.

Carrying out a similar process for the linear down metric, we note a number of key findings illustrated by the following terms:

social refers to the Bush administration's attempt to create a legacy issue around a considerable change to the Social Security system in the US.

gitmo refers to the holding prison in Guantanamo Bay for prisoners of war captured in Bush's War on Terror.

downing refers to a scandal involving the ' Downing Street Memo' - connecting both the British and US administration with fabricated intelligence regarding the invasion of Iraq.

prisoners refers to both the Guantanamo prison in Cuba and the Abu Grahib prison in Afghanistan.

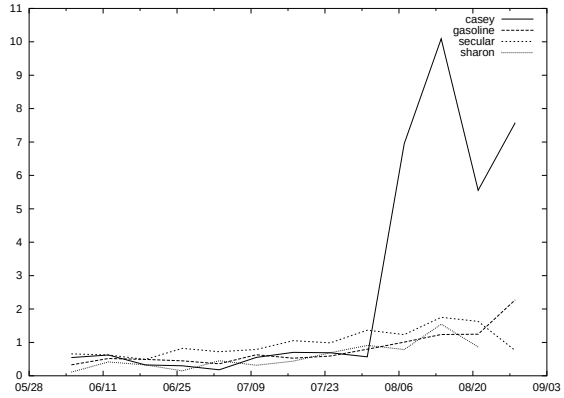

Figure 2: A sample of terms mined using the linear up metric.

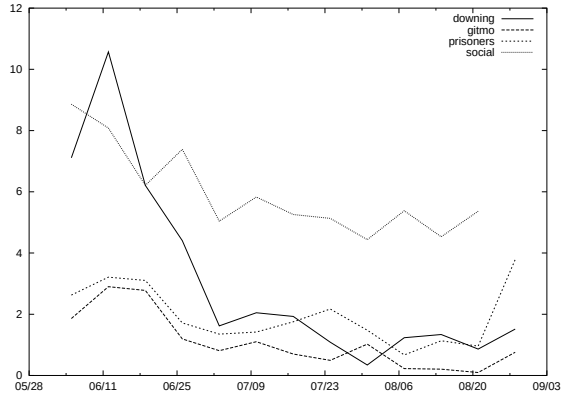

Figure 3: A sample of terms mined using the linear down metric.

In summary, the issues which are seeing a decline in discussion in the blogosphere include those related to specific actions and consequences of the War on Terror as well as Bush's controversial Social Security reform agenda.

By comparison, Table 2 shows the top 20 keywords by rank and by frequency mined in an atemporal manner as described in (TH03). It is interesting to note that these sets have quite a different flavour and certainly cannot qualify the movement of the terms - a factor which greatly contributes to determining what actions if any ought to be taken by those analysing closely this sort of data.

Table 3 shows the top 10 tokens ranked by the max burst metric. The terms are dominated by those refering to the hurricane Katrina disaster in New Orleans. It is interesting to note that during that time period there was a significant increase in posts about George Bush specifically associated with the disaster (see Figure 1). The hurricane event follows a typical pattern - an initial acute spike followed by rapid decay.

Even though the time series are normalized against that background, the Katrina concept is well captured by this measurement.

Rank	Frequency
bb	war
myspace.com	iraq
mystarlinks.com	times
rearrange	u.s
july	john
rove	support
war	military
london	case
hairy	court
nude	july
polska	press
wilson	free
outdoors	democrats
stars	june
iraq	information
remiel	supreme
jolie	flight
girls	senate
chics	women
sokolova	troops

Table 2: The top 20 keywords by rank and by frequency.

Term	max burst
katrina	12.28
levees	11.71
fema	11.09
orleans	10.85
hurricane	10.78
superdome	10.62
levee	10.34
underprivileged	9.97
looters	9.90
louisiana	9.61

Table 3: The top 10 terms for the max burst metric.

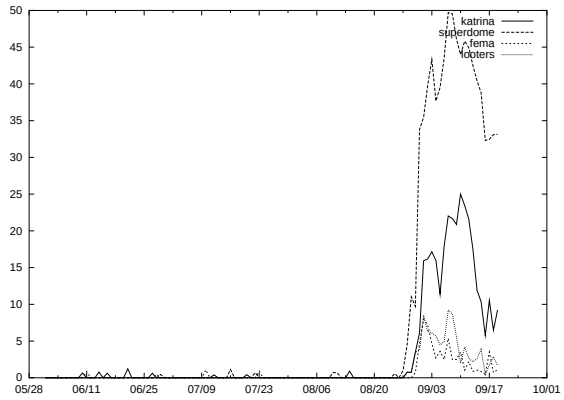

Figure 4: Terms mined with the max burst metric plotted in daily granularity.

Term	F-Burst
rita	7.35
pyongyang	6.95
mugging	6.92
newsman	6.51
chump	6.45
inflationary	6.42
breakthrough	6.12
prizes	6.09
no-bid	5.93
fiscally	5.88

Figure 4 shows the time series for a sample of these terms. The series have been plotted in day granularity (not week granularity) to give more texture to the story.

Table shows the top 10 terms as ranked by the final burst metric. Here we can see a number of other issues coming to the fore, including hurricane Rita and a story about bilateral talks with Pyongyang.

Further Work

The temporal models presented in this paper illustrate a class of analyses that may be carried out on text data which has a temporal dimension. The models in themselves are trivial, but capture intuitive and useful patterns readily applicable in a number of text mining scenarios.

This paper has made no attempt to provide a strong evaluation of the methods presented. This paper is describing the application of standard techniques to a textual data set in the context of a particular application - mining online data for corporate tasks such as brand monitoring, product feedback, etc. Evaluation, therefore, should be in terms of how this method improves some task. For example, does trend mining of this sort improve the precision of alerting mechanisms?

The system described in this paper can be considered as a keyword discovery mechanism: words are ranked according to some metric relating to temporal patterns. It would be interesting to explore the space of phrase mining as well as

association mining using temporal features.

Conclusions

This paper has presented an application of time series analysis to mining weblog data.

References

Fernando Diaz and Rosie Jones. Using temporal profiles of queries for precision prediction. In *Proceedings of SIGIR, International Conference on Information Retrieval 2004*, 2004.

Takashi Tomokiyo and Matthew Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, 2003.