Searching for Events in the Blogosphere

Manolis Platakis
platakis@di.uoa.gr

Dimitrios Kotsakos
d.kotsakos@di.uoa.gr

Dimitrios Gunopulos
dg@di.uoa.gr

Dept. of Informatics and Telecommunications,
National and Kapodistrian University of Athens, Greece

ABSTRACT

Over the last few years, blogs (web logs) have gained massive popularity and have become one of the most influential web social media in our times. Every blog post in the Blogosphere has a well defined timestamp, which is not taken into account by search engines. By conducting research regarding this feature of the Blogosphere, we can attempt to discover bursty terms and correlations between them during a time interval. We apply Kleinberg's automaton on extracted titles of blog posts to discover bursty terms, we introduce a novel representation of a term's burstiness evolution called *State Series* and we employ a Euclidean-based distance metric to discover potential correlations between terms without taking into account their context. We evaluate the results trying to match them with real life events. Finally, we propose some ideas for further evaluation techniques and future research in the field.

Categories and Subject Descriptors

H.4.m. [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement, Experimentation

Keywords

Blogs, social media, burst analysis, hot topics, keyword correlation, information retrieval, text mining

1. INTRODUCTION

Everybody reads blogs. Almost everybody maintains one. Wikipedia defines a blog as a website, usually maintained by an individual, with regular entries of commentary, descriptions of events, or other material such as graphics or video. Over the last few years, blogs (web logs) have gained massive popularity and have become one of the most influential web social media in our times. Anyone with an internet connection can create his own blog for free, using web platforms developed for this specific reason (e.g. blogger.com, wordpress.com etc.). According to blog search engine Technorati.com there are over 175,000 new blogs every day, 1.6 million new posts per day and over 113 million blogs (not including millions of non-English blogs) exist today. The huge growth of blogging provides a wealth of information waiting to be extracted.

Blog analysis and searching in blogs introduces new challenges for research in information retrieval because blogs' contents have a very specific characteristic not present in traditional web content: a *timestamp* exists in every blog post. Every blog post in the Blogosphere has a well defined value in the temporal axis. Traditional blogs' search engines don't take into account the temporal dimension and treat the blogs as plain web content; or just pay attention to the category tags that may accompany a post. By taking into consideration the timestamp of each blog post we can try to detect the period in which the popularity of a specific keyword increases or decreases. Such functionality is important because it allows us to gauge the users' interests related to a specific topic over time.

Our contribution: In this paper we develop a technique to address the problem of identifying events in the Blogosphere. In our technique we apply Kleinberg's automaton [2] on extracted titles of blog posts to discover bursty terms, we introduce a novel representation of a term's burstiness evolution called *State Series* and we employ a Euclidean-based distance in order to discover potential correlations between terms without taking into account their context.

Related work: As the number and size of large time-stamped collections increases this problem becomes more and more important [1], resulting in an evolution clearly presented in [3]. Due to space constraints we refer to [4] for a more extensive review of the related work.

2. OUR APPROACH

We search for events in the Blogosphere. We define an *event* in the Blogosphere as a small subset of keywords able to describe one or more real life events that occurred during the period of study. To discover them we try to identify *correlated* bursty terms, meaning bursty terms whose burstiness exhibits a similar behavior in the temporal axis. A *burst* is marked whenever the popularity of a specific keyword dramatically and unexpectedly increases. Doing so, we omit taking into account a keyword's possible co-existence with another keyword in the same title. Ignoring all those keyword pairs enables us to gain significant computational time and to search for conceptually correlated keywords although they may not appear in the same document (e.g. separately used synonyms).

In order to identify bursty terms, meaning specific words whose appearances increase radically in short periods of time in comparison to the long period we study, we use the technique proposed by Kleinberg [2] as decribed in [4]. Afterwards we evaluate the accuracy of the bursty terms by trying to match them with real life events that took place in the bursty period of time. A certain event is formed by a group of correlated terms. As the popularity of a specific topic diminishes, this group ceases to exist. We try to obtain keywords' correlations, in order to automatically identify such groups. We address this problem, assuming that related keywords produce similar activity as far as burstiness is concerned. A mechanism for burstiness representation of a term *t* called *State Series* is introduced and is defined as follows:

$$SS_t = (s_{t1}, \dots , s_{ti}, \dots ,s_{tn}),$$

where s_{ti} represents the burstiness state of term t at timestamp i, produced by the automaton. Figure 1 compares the frequency curve of the term *indiana* to the corresponding SS, proving that the latter is a satisfactory representation. Furthermore we employ

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

a Euclidean-based distance metric to calculate the dissimilarity between the *SSs* of two different terms. Finally, we obtain events by accumulating the 5 Nearest Neighbors for each keyword, assuming that 5 terms can adequately describe an event. Last but not least we evaluate these events by trying to pair this topic with a real life event that took place in the period of study.

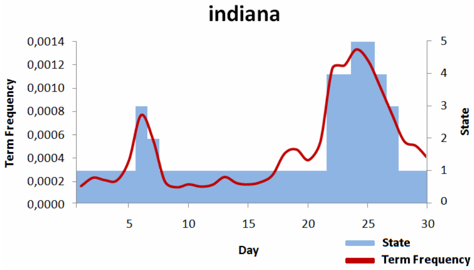

Figure 1. Frequency curve and SS of the keyword "indiana"

3. EXPERIMENTAL EVALUATION

Our experimental evaluation shows that blog posts' titles prove sufficient to mine the underlying bursts. Not using the whole body of each blog post reduces the total computational time required. Therefore, we extend this approach to search for events through the burstiness pattern of keywords appearing in blog posts' titles.

Data description: We experimented on posts from millions of blogs around the web's free blog hosts (e.g. blogger.com, wordpress.com, livejournal.com etc.) After some pre-processing of our initial dataset we ended up with 11,198,076 titles containing 38,814 different keywords with various appearances during the period May 1 – May 30, 2008.

We used an *n*-state automaton, incrementing *n* and monitoring the percentage of the terms with altered 5-NNs in comparison to the results of the *n*-1-state automaton. As shown in Figure 2, the greater state value that could be reached was 13. The ratio of the exponential rate of the automaton's each subsequent state to the rate of the previous state was picked to be 1.3, after several experimental trials. This value provides us with increased diversity in the state series. The automaton identified 21.53% of the terms as bursty.

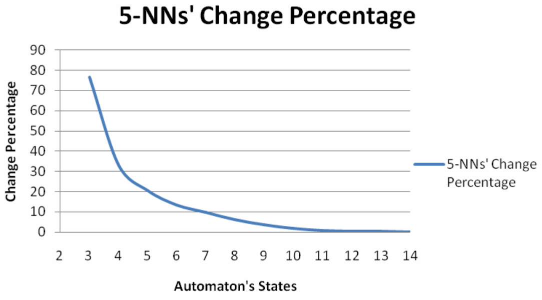

Figure 2. Changing the number of states

Semantic evaluation: A visualized example of identifying bursty keyword correlations using the SS similiarity is depicted in Figure 3, where the SSs for the 3-NNs of the term Indiana are shown. The terms i*ndiana, jones, crystal* and *skull* appear in the results as bursty ones. While trying to evaluate the accuracy of this result, we found out that on May 22nd 2008 the movie "*Indiana Jones and the Kingdom of the Crystal Skull*" was released. Additional results shown in Table 1 add to the assumption that events can be mined through the extraction of state series.

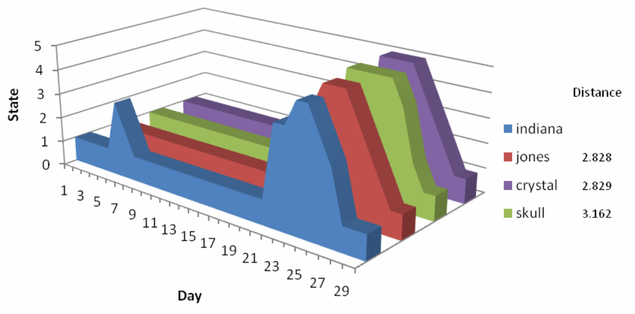

Figure 3. A comparison of the SS of related keywords

As seen in Figure 3, in this case the proposed method came out to be resistant to effects of other bursts of a term, that seem to be irrelevant to the event being described by the 5-NNs; *indiana* exhibits two bursts during May, one lasting from 6th to 7th day and one from 22nd to 27th day, but the former one does not affect the high similarity between *indiana* and the other three terms.

Table 1. Semantic evaluation results

Term	*5-NNs*	*Burst Intervals*
pharaoh	{physique,feminine, akhenaten,liver,transplant}	(2-3)
liver	{transplant,marijuana,wig, feminine,physique}	(2-3)
myanmar	{burma,burmese, appreciation,chait,brutality}	(5-14)
cialis	{tadalafil,trent, prescription,pharmacy, impotence}	(5-6) (8-9) (19-20)
indiana	{jones,crystal,kingdom, skull,islander}	(6-7) (22-27)

4. ONGOING AND FUTURE WORK

Our results regarding events could be improved if we were to examine burstiness similarity in sub-intervals during a longer period of time; a direction which we are eager to explore in the near future. We plan to evaluate the precision of our method by calculating the percentage of the NNs that actually co-exist in the same documents with the examined term. This work has been partially supported by the SemsorGrid4Env EC project.

5. REFERENCES

[1] Bansal, N., Koudas, N., BlogScope: A System for Online Analysis of High Volume Text Streams. VLDB 2007: 1410-1413.

[2] Kleinberg, J. M., (2003), Bursty and hierarchical structure in streams, Data mining and Knowledge Discovery, 7(4): 373-397.

[3] Kumar, R., Novak, J., Raghavan, P.,Tomkins, A. On the Bursty Evolution of Blogspace. World Wide Web 8(2): 159-178 (2005).

[4] Platakis, M., Kotsakos, D., Gunopulos, D. Discovering Hot Topics in the Blogosphere. In Proc. of the 2nd Panhellenic Scientific Student Conference on Informatics, Related Technologies and Applications EUREKA 2008, pp. 122-132.