Searching for Events in the Blogosphere

Manolis Platakis*, Dimitrios Kotsakos*, Dimitrios Gunopulos†

*Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece
†Department of Computer Science and Engineering, University of California, Riverside

Motivation: People blog + Things happen = People blog about things that happen

How can we identify such events?

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.

Our Approach:

- Each blog post has a timestamp
- Popularity of a term may increase in time, marking a burst
- We extract bursty terms from blog post titles using Kleinberg’s automaton
- We represent burstiness behavior introducing State Series:
  \[ SS_t = (s_{t1}, \ldots, s_{tn}) \]
  \( s_i \) is the burstiness state of term \( t \) at timestamp \( i \), produced by the automaton
- We assume:
  Similar SSs \( \rightarrow \) Potential term correlation
- We employ a Euclidean-based distance metric

Experimental Results:

Description: We experimented on posts from millions of blogs. Our dataset consisted of 11,198,076 blog post titles which were posted on the web during the period May 1 – May 30, 2008. We obtain events by accumulating each term’s 5-NNs and map them to a real life event.

Query term: “pharaoh”

5-NNs: physique, feminine, akhenaten, liver, transplant

Query term: “indiana”

5-NNs: jones, crystal, kingdom, skull, islander
Event: “Indiana Jones and the Kingdom of the Crystal Skull was released worldwide on May 22, 2008.”

Query term: “cialis”

5-NNs: tadalafil, trent, prescription, pharmacy, impotence
Event: “Tadalafil is the main ingredient of the prescription drug Cialis administered orally to treat male impotence and can easily be found in any pharmacy.”

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 the figure, the greater burstiness state value reached was 13.

As shown in the figure the proposed method came out to be resistant to effects of other bursts of a term, irrelevant to the event being described by the 5-NNs.

Ongoing and future work:

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. We are also eager to examine burstiness similarity in sub-intervals during a longer period of time. We are currently exploiting other mechanisms, based on discrepancy theory, to identify bursty intervals [2]. Moreover, we look forward to exploiting tags and multimedia metadata that may accompany a blog post.

References: