BlogScope: Spatio-temporal Analysis of the Blogosphere

Nilesh Bansal
Department of Computer Science
University of Toronto
nilesh@cs.toronto.edu

Nick Koudas
Department of Computer Science
University of Toronto
koudas@cs.toronto.edu

ABSTRACT

We present BlogScope (www.blogscope.net), a system for analyzing the Blogosphere. BlogScope is an information discovery and text analysis system that offers a set of unique features. Such features include, spatio-temporal analysis of blogs, flexible navigation of the Blogosphere through information bursts, keyword correlations and burst synopsis, as well as enhanced ranking functions for improved query answer relevance. We describe the system, its design and the features of the current version of BlogScope.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Algorithms

Keywords: Text Analysis, Blogs, Trends, Visualization, Information Discovery

1. INTRODUCTION

Blogs have been proliferating over the last couple of years. It is estimated that the size of the Blogosphere in August 2006 was two orders of magnitude larger than three years ago [3] . Blogging activity includes personal diaries, traveling experiences, opinions (about products, events, people, music groups, or businesses), howto guides, and politics. Collecting, monitoring and analyzing information on blogs can provide key insights on 'public opinion' on a variety of topics, such as products, political views, or entertainment. It can also be a source of competitive intelligence information and market trends [1]. As a result, techniques that aid the collection, analysis, mining and efficient querying of blogs are important and this trend is expected to persist, given the growing popularity of blogs.

The information in blogs and its dynamics differ from the traditional web content. Significant differences include: (1) blog posts have a time of creation adding a temporal dimension, (2) blog posts may trigger additional posts by the same or other bloggers leading to a discussion in the Blogosphere, and (3) blog posts can be easily associated with a geographical location which is the same as the location of the author[1]. We introduce BlogScope, a system with enhanced analysis capabilities (well beyond keyword search) for blogs[2].

[1]Traditional websites, e.g., en.wikipedia.org or www.yahoo.com, do not have a well defined geographical location.

[2]Although we confine our discussion on blogs we believe that much of our discussion is pertinent to all temporally ordered streaming text sources like mailing lists, forums, news groups etc..

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

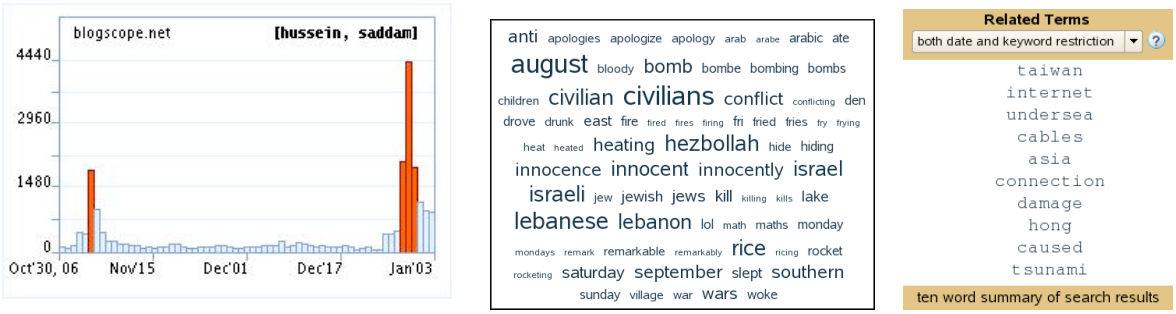
Figure 1: GeoSearch for the query *iphone*. Black dots on the map represent regions where bloggers are writing about the searched query.

2. ANALYZING THE BLOGOSPHERE

The analysis paradigm that BlogScope facilitates is segmented in four steps. BlogScope identifies *what* is 'interesting', *when* it was 'interesting', *why* it is 'interesting', and *where* it is 'interesting'. On its front page, BlogScope displays a list of hot keywords. Such keywords are computed daily from the actual content of blog posts. Based on this list, a user can formulate a query to seek for relevant blog posts. The traditional text query interface is also supported to identify posts relevant to a query, in case one is seeking for specific information. Once the keywords of interest are identified, a query is formed and relevant blog posts are retrieved. The next question BlogScope aids to answer is when it was interesting. To answer this question, BlogScope plots the popularity of the query keywords in blog posts, as a function of time, and identifies and marks interesting temporal regions as bursts in the keyword popularity. The third step of the analysis is to investigate why it is interesting. Correlated keywords (intuitively defined as keywords closely related to the keyword query at the specified temporal interval) are automatically displayed by BlogScope. Such keywords aim to provide explanations or provide insights as to why the keyword experiences a surge in its popularity. Based on these keywords, one can refine the search and drill down in the temporal dimension towards a more focused subset of blog posts. The final step is to identify where it is interesting. BlogScope associates with each blog its geographical coordinates. This information is used to annotate the world map with regions where bloggers are writing about the searched query.

Figure 1 and 2 provide example screenshots from BlogScope demonstrating various features. It must be noted that all the analysis is performed on the actual textual content of blog posts, and not on *tags* because: (1) tagging requires manual effort, (2) most blogs posts are not tagged, and (3) a few tags can not accurately capture complete information present in a post.

Popularity and Bursts: The popularity curve for a keyword (or set of keywords) displays how often the specified keywords are mentioned in the Blogosphere as a function to time. Such a curve



(a) (b) (c)

Figure 2: (a) Popularity curve for the keywords *saddam hussein*. Saddam was convicted on November 5 2006, and was executed on December 30 2006. Regions marked in red indicate bursts. (b) Hot keywords cloud tag for July 30 2006. (c) Correlations for keyword *earthquake* for December 27 2006, soon after an undersea earthquake in taiwan resulting in disruption of internet services.

and its fluctuation can provide insight regarding the keyword popularity evolution over time. An unexpected increase in the popularity of a keyword indicates the occurrence of an event. Such events are captured by BlogScope as *keyword bursts*, and are marked in red on the popularity curve. The popularity of different keywords can be compared for analysis purposes.

Correlations: Information in the Blogosphere is highly dynamic in nature. As topics evolve, keywords align together to form stories; and as topics recede, these keyword clusters dissolve. This formation and dissolution of clusters of keywords is captured by BlogScope in the form of correlations. With every search, a list of keywords in blog posts most closely related to the search query keywords is displayed. Roughly speaking, such keywords are those that most frequently co-occur with the searched query terms (weighted by their *idf*). They aim to provide insight regarding the posts relevant to the searched queries.

Geo Search: With every post indexed by BlogScope, a city, state and country is maintained. This is done by extracting location string from author profiles, and utilizing approximate match technology (e.g. [2]) against lists of known cities. As a result, BlogScope can display the distribution of the posts on a map, or restrict the search to select regions of the world.

Hot Keywords: On its front page BlogScope displays a list of *hot keywords* for that day in the form of a cloud tag. BlogScope uses a measure of 'interestingness' for keywords and ranks all keywords for a day according to this measure. Interesting does not necessarily refer to popular. For example, keywords that exhibit sudden change in their popularities are more interesting. In a few occasions, Blogscope tracked popular keywords that corresponded to events that have not made mainstream news media. For example the term *math* was highly popular on the week of August 7 2006 in the Blogosphere as reported by BlogScope. The event corresponded to the news about the Poincare conjecture proof by Grigory Perelman. New York Times had an article on this on August 15 2006.

Burst Synopsis: In order to aid information discovery, BlogScope incorporates features that aim to explain events related to a search query. In particular once a burst is identified for a query string, BlogScope can generate a *synopsis set* for it. The semantics associated with the burst synopsis set for a query q is that it is the maximal set of keywords correlated with q that exhibit a bursty behavior in the associated popularity curves. For example, consider the query *saddam* on December 30 2006 (the day when he was executed). BlogScope generates the set *saddam, hussein, hanging, execution, dictator, iraq, brutal* as its burst synopsis, which includes all keywords that exhibit a burst on the selected day in connection to the query, serving as a description of the event.

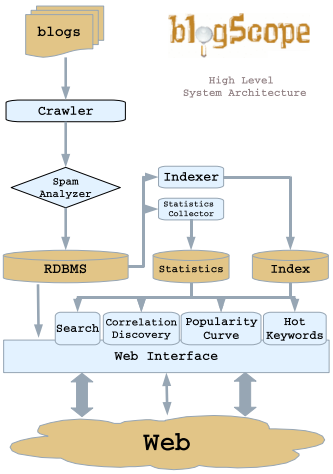


Figure 3: High level system architecture of BlogScope.

3. CONCLUSIONS AND FUTURE WORK

We presented BlogScope, a text analysis system suitable for temporally ordered streaming text, currently applied to the analysis of the Blogosphere. At the time of this writing, it was tracking over 4.5 million blogs with around 40 million posts in its database. Everyday around 100 thousand new posts are added to the system. It is extremely important, given the analysis the system conducts, for the techniques employed to be computationally efficient in order to scale at this level. We have therefore developed effective and efficient algorithms for burst identification, discovering correlated terms, mining hot keywords, and burst synopsis generation. Figure 3 presents the overall architecture of the system highlighting its main components. We plan to continue enhancing BlogScope with several features to improve navigation, information discovery and performance.

4. REFERENCES

[1] D. Gruhl, R. V. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *SIGKDD*. ACM, 2005.

[2] A. Chandel, O. Hassanzadeh, N. Koudas, M. Sadoghi, and D. Srivastava Benchmarking Declarative Approximate Selection Predicates In *SIGMOD*, 2007.

[3] State of the Blogosphere - aug 2006. *http://www.sifry.com/alerts/archives/000436.html*.