BlogTrackers: A Tool for Sociologists to Track and Analyze Blogosphere

Nitin Agarwal[*], Shamanth Kumar[*], Huan Liu[*], Mark Woodward[+]

[*]Computer Science & Engineering, School of Computing & Informatics

[+]Department of Religious Studies, College of Liberal Arts and Science

Arizona State University, Tempe, AZ 85281

{nitin.agarwal.2, shamanth.kumar, huan.liu, mark.woodward}@asu.edu

Abstract

We present a tool BlogTrackers, which assists sociologists to track and analyze blogs of particular interests by designing and integrating unique features. We present an overview of BlogTrackers, illustrate its functions of various components of BlogTrackers, and outline future work for expansion in meeting the growing needs of sociologists.

Introduction

Blogosphere, the network of blogs, is growing at a phenomenal rate. Technorati[1] has indexed around 133 million blog records. Sociologists are interested in studying the blogosphere for tracking socio-behavioral patterns, identifying the influential people in the region of interest and tracking interesting activities. They often have to eyeball the sites for useful information. Given a gamut of interests in the blogosphere, this can be a tedious and time consuming task. Through this user-oriented application, we propose to alleviate this problem by assisting them in effectively tracking and analyzing blogosphere. BlogTrackers grants sociologists the freedom to choose the blog sites they wish to analyze, observe interesting events and patterns with the flexibility of drilling-in. The tool consists of a number of analyzing and crawling modules and is a convenient alternative to eyeballing the blog sites and concentrate efforts on further analysis.

Most tools are generic in nature and cannot be directly used by sociologists and others with specific needs. BlogTrackers is particularly designed for their needs that can perform both data collection and provide convenient visualizing tools to analyze the data. Table 1 presents a comparison of BlogTrackers with some of the existing tools in the domain. Although, sites like Technorati and BlogPulse provide features similar to our tool, they cannot be directly used. BlogTrackers combines them in a unique manner to maximize the analytical capability of the individual techniques. Apart from these tools there is also some generic visualization software that do not target blogs per se but can be used to do some analysis on the blog data. Pajek is a visualization tool that can be used to visualize the network data in various ways. IBM's ManyEyes is another interesting project on generic visualizations but suffers from scalability issues. The Prefuse visualization toolkit contains a set of unique visualizations for the data.

[1] http://technorati.com/blogging/state-of-the-blogosphere/

BlogTrackers

BlogTrackers is a Java based desktop application that provides a unified platform for the user to crawl and analyze blog data. It grants the user, the freedom to choose the data of interest and helps in effectively analyzing it. The data is stored in a relational database. Currently we are tracking 10 different data sources like Twitter, Engadget, The Unofficial Apple Weblog (TUAW), LiveJournal, etc.

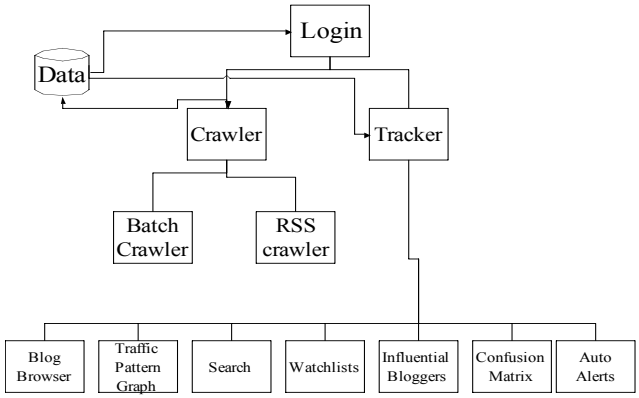

Figure 1.BlogTrackers component diagram.

Crawler

BlogTrackers offers two types of crawlers to the user. The *batch crawler* crawls the websites from scratch and stores it in a database through HTML scraping using regular expressions to parse data from the HTML files. The *RSS* (Really Simple Syndicate) *crawler* on the other hand incrementally crawls the websites by retrieving information from their feeds. RSS crawler can be scheduled to run automatically and update the database.

Tracker

The tracker component provides the user with a set of tools to analyze the data. The blog site to be used for analysis can be chosen by the user. The following are the 4 major tools in BlogTrackers.

	Source Selection	Data crawler	Influential Bloggers	Watch Lists/Alerts	Blog Browser	Traffic Trends	Search	Conversation tracker	Keyword Trends	Tag Clouds
BlogTrackers	✓	✓	✓	✓	✓	✓	✓	×	×	✓
Technorati	✓	✓	×	✓	✓	×	✓	×	✓	✓
BlogPulse	✓	✓	×	×	×	×	✓	✓	✓	✓
BlogScope [Bansal 2007]	✓	✓	×	×	✓	✓	✓	×	✓	✓
GTD Explorer [Lee 2008]	✓	--	×	×	×	✓	✓	×	✓	×
IceRocket	✓	✓	×	×	×	×	✓	×	✓	✓
Google Blog Search	✓	✓	×	✓	✓	×	✓	×	×	×

Table 1. Comparison of BlogTrackers with other tools

1) Blog Analysis. BlogTrackers contains a blog browser that can be used to individually analyze the blog posts within a time period. The time window can be adjusted as required. Entire blog posts are indexed for better viewing experience. Another tool for blog analysis is *Term Frequency Analyzer*, which shows the tag cloud (a visualization which highlights the terms by their frequency by varying the font) for all the blog posts in a given time period. This tool can be used to identify key terms associated with the blog posts during a particular time period. A *traffic pattern graph* can also be generated for a particular time period (Figure 2). The bar graph shows the traffic bursts depending on the granularity chosen (daily, weekly, monthly, or yearly). The bursts can be individually analyzed to observe the blog posts and the tag cloud for that period.

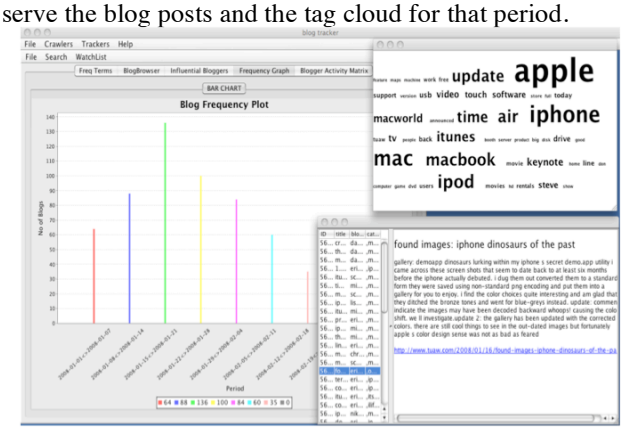

Figure 2. Traffic Pattern Graph showing the Tag Cloud.

2) Blogger Analysis. BlogTrackers can be used to search for influential bloggers at a blog site. The influential bloggers are generated as described in [Agarwal et al. 2008]. It is also possible for a user to drill-in and look at the tags and the blog posts of the influential bloggers. Bloggers can be classified based on their activity and influence into different categories like "Active-Influential", "Inactive-Influential", "Active-Non Influential", and "Inactive-Non Influential". These categories can be visualized as a confusion matrix as in Figure 3.

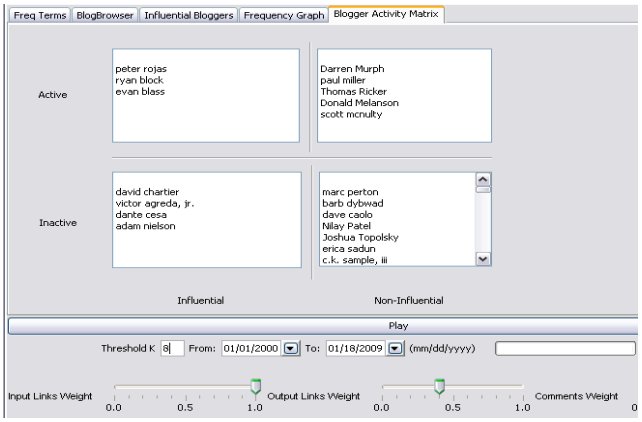

Figure 3. Analyzing the Blogger Categories.

3) Search. The blog sites are crawled on a daily basis and the posts are stored and indexed using Lucene. The index is automatically updated and can be used to search the blog posts for specific queries.

4) Watchlists/Alerts. BlogTrackers offers a convenient notification system to the users. A user can specify terms in the watchlist. The user is then notified by e-mail if any new post contains that word. The user can also choose the terms from a system-generated list based on popularity.

Future Work and Conclusion

The features described above forms a part of the solution to the problem. In future we intend to add more features and improve usability. Some of these features include sentiment analysis, blog clustering, identify hot topics, keyword-trend graphs, and conversation tracking. We also plan to release the API for BlogTrackers to share data (under GNU's GPL) and various analysis features for a community wide collaborative development and research. The blogosphere is growing every year at an unprecedented rate so having as much data as possible to analyze is crucial therefore we plan to expand the data repository and include more blogs to better serve sociologists needs.

Acknowledgement

This project is, in part, sponsored by ONR grants N000140810477 and N00014-09-1-0165. We also appreciate the valuable help of Alan Zheng Zhao.

References

Agarwal, N.; Liu, H.; Tang, L.; and Yu, P. 2008. Identifying Influential Bloggers in a Community. In *Proc.of the Intl.Conf.on Web Search and Data Mining*.

Bansal, N., and Koudas, N. 2007. Searching the Blogosphere. In *Proc. of the Intl.Workshop on the Web and Databases*.

Lee, J. 2008. Exploring Global Terrorism Data. *ACM Crossroads* 15(2):7-14.