BlogHarvest: Blog Mining and Search Framework

Mukul Joshi*

Great Software Laboratory
Pune
India

mukul.joshi@gs-lab.com

Nikhil Belsare

Tech Mahindra Ltd.
Pune
India

nikhil.belsare@techmahindra.com

Abstract

Beyond serving as online diaries, weblogs have evolved into complex social structures. Blogging software allows users to publish opinions on any topic without any constraints on the predefined schema. Analysis of linkage between blogs has indicated that community forming in blogosphere is not a random process but is a result of shared interests binding bloggers together. Learning, analysis and usage of the user's interest and social linkage from the blog is therefore necessary to provide useful search faculty on the blogosphere to bloggers and revenue generation opportunities like advertising to the blog service providers.

In this paper, we demonstrate BlogHarvest which is a blog mining and search framework that extracts the interests of the blogger, finds and recommends blogs with similar topics and provides blog oriented search functionality. BlogHarvest uses classification, linkage & topic similarity based clustering and POS tagging based opinion mining for providing these features. Novel search interface is built to provide related blogs for queries along with the usual result ranking. Association rules found from POS tags are used to get the context of search for providing query expansion to get targeted results. By crawling the blogosphere and extract & index blog posts and linkage metadata; we have analyzed around 50000 blogs to tune our algorithms.

1. Introduction

A growing number of blogs are being published on the Internet. As per the data published by Technorati Inc. [13] over 70,000 new weblogs get created each day and over 50,000 postings take place per hour but co-relating the information published on these blogs is the next big challenge.

Example:
User searches 'AJAX' on a blog search engine and gets only the results that contain the query term 'AJAX'. But the user may like to reach a blog rich in information on AJAX. The user may perhaps be interested only in the disadvantages of AJAX and hence may only look for search results carrying negative opinions. Current search engines fail prey towards these needs. [4, 12]
For bloggers and frequent blog readers, it is virtually impossible to keep track of the growing blogosphere and hence a service recommending the blogs matching their interests will seek high value. Classifying blogs to get the blogger's profile and identifying blogger clusters exploiting bloggers social network is therefore an important task required to provide value-rich services to the blog search engine users and the bloggers themselves. Though there are many research prototypes [3, 7] that try study and apply these data mining techniques to blog data, we haven't come across any end-to-end framework that combines these various aspects together targeting the needs of bloggers and searchers. With BlogHarvest we bridge this gap and also provide some novel alternative techniques for solving the various blog mining and search sub-tasks.

2. BlogHarvest's features

BlogHarvest has following features:

1. Analysis of blogs at content level.
2. Analysis of semantics of the information published on the blog.
3. Exploting blogger network for search and recommendation
4. Formation of clusters of bloggers having similar interests
5. Recommending users blogs of their interests

International Conference on Management of Data
COMAD 2006, Delhi, India, December 14-16, 2006

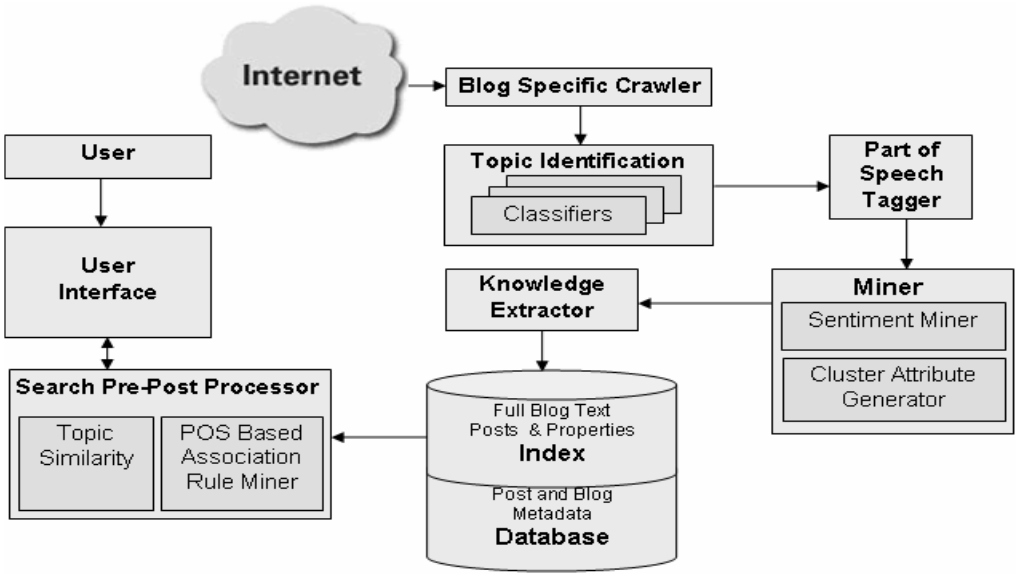

Figure 1: System Architecture

6. Building a profile of blogger based on the content published in the blog.

Analysis and search in BlogHarvest differ from the other similar efforts by incorporating the topical and linguistic attributes of the posts and structure of a blog than just concentrating on the pure text.

2. System architecture

2.1 Blog host specific crawler

Every blog host has a different structure for all the blogs published on it. The crawler extracts posts from blogs and attributes of those posts (Title, Timestamp, Body, Permalink etc.). Since our analysis relies on the blogger network analysis, we have separate crawlers per blog host to extract this information. Scheduling the crawlers for capturing the updates in the blogs efficiently is currently not handled in our framework

2.2 Topic Identification

Topic Identification is used to determine what the user frequently writes about. Topic Identification is based on the Naïve Bayes classification using rainbow. [10]

The classifier is trained with the DMOZ[2] rdf dump. It is also trained with anchor text of the links pointing to the web pages found in DMOZ dump. A base model is built with 14 categories in it. A hierarchical classification scheme is used: text from the posts is first submitted to the base classifier to get base class and then the same text is submitted to the specialized classifier of that class to get super fined classes. Specialized classifiers are trained for few base categories such as Books, Movies and Music etc. Three most dominant classes of the post are only considered based on the probability of each class. However if the highest class probability is more than twice that of the next higher probability then we discard the other two classes.

2.3 Part-of-speech Tagging

In Part of speech tagging every word in the text is tagged with its part of speech. For example, adjective excellent will be tagged with _JJ to become excellent_JJ. This information is then used for sentiment analysis and formation of association rules for query expansion. QTag [9] – an English language Part-of-speech tagger is used for tagging the text.

2.4 Mining the blogs

Sentiment Miner: Sentiment analysis deals with predicting the sense of the post text. For example, if user has written a movie review then sentiment analyser tries to predict whether user is criticising the movie or he is praising the movie.

One approach is to have a seed list of adjectives of known orientation. The list is prepared manually with the help of WordNet[15]. Then at sentence level

- If more adjectives of certain polarity, the sentence has that polarity orientation
- If there is a negation word (not, neither etc.) left to an adjective, the current polarity is negated. [8]

Another approach is to train classifier with movie review data such as IMDB[5] review repository. Reviews with low rating are used to train negative classifier and reviews with higher rating are used to train positive classifier. Then these classifiers are used to determine sentiments. [11] We are currently evaluating this approach.

Knowledge Extraction: Out links in the posts are extracted and classified to increase accuracy of prediction of user's interests. Further the web sites frequently visited by blogger are also determined.

Clustering: Users with similar interests should be clustered together. The clustering is done based on three attributes.

- Weight for every class match – For every blog, the topic probabilities for entire blog are calculated by adding up and normalizing class probabilities of each posts. Then similarity between given two blogs, A and B is calculated by root mean square for every match.

$$\sqrt{p(a) * p(b)}$$

  Where p(a) is probability of some topic for user A and p(b) is probability of the same topic for user B
- Weight for out link intersection – for example, if blog A and blog B contain links to the same web page then we can say that it may make the blogs as candidates to be put in same clusters
- Weights for linkage distance – above two attributes are then normalized by the linkage distance found by counting the number of intermediate blogs in the bloggers network.

These attributes are then used by K-nearest neighbours clustering algorithm [6] to form clusters. As mentioned previously, BlogHarvest can recommend new blogs to the bloggers. This is done by clustering the bloggers having similar interests from the crawled blogosphere.

2.5 Indexing and meta data

Posts extracted from the blogs are indexed using Apache Lucene [1] to provide search. Two separate indexes are built: a common index that stores posts from blogs and a per user index to provide search over particular blog.
Blogger network information and other meta-data are stored in mysql database.

2.6 Search pre-post processor and User Interface

Topic Similarity: Along with the search results users are also provided with a list of blogs that are relevant to the topic of the search query i.e. users are recommended different blogs that they may want to read. Topic that occurs maximum number of times in the search results is found and blogs that have high scores for that topic are then declared as relevant blogs.

Part-of-speech based Association Rule miner: Targetting the result pages for the search intent is many a times a daunting task given very high number of matching results for a query. We are exploiting the POS tagged text to provide search interface for query expansion. Consider an example of a query "Newton". There are various dimensions to the query which the results may talk about. For example, Newton's contribution to Physics, Mathematics, Calculus, Geometry or even his absent mindedness, user may be willing to find results for one of these and hence grouping the results based on this dimensions is one way by which clustered search engines try to tackle this problem. [14]
However we handle this problem by displaying frequent Nouns, Verbs and Adjectives associated with the query as the 3 dimensions over the search space lattice.

User Interface: A novel interface is provided to the end user for searching the blogs. The interface also provides visualization of network and clusters of bloggers.

3. Description of the demo

Use Case: Search

We will demonstrate how BlogHarvest provides a unique search experience for user. Figure 2 shows a snapshot of search interface in action. As can be seen in the figure, the user is provided with list of blogs that are relevant to the query in the right pane. The cached versions of result posts are also made available to the user. Users are provided with the polarity of the posts along with the associated confidence.

Use Case: Blogger's Profile

We will demonstrate that BlogHarvest generates profiles of users based on the knowledge extracted. Figure 3 shows a sample profile. Profile contains i) Blogger's interests, ii) Sites frequently visited by the blogger, iii) Visualization of fellow blogger network of the blogger, iv) Entire blog can be syndicated as RSS v) Left pane displays user's opinions about various topics and the related keywords.

4. Conclusion

Blogs are a source of enormous information. For a user it is very hard to get the relevant information from this huge network. BlogHarvest provides means to correlate the information found on different blogs in this network andprovides search facility to user. As a part of future work, we plan to provide RSS feeds of blogs of user's interest.

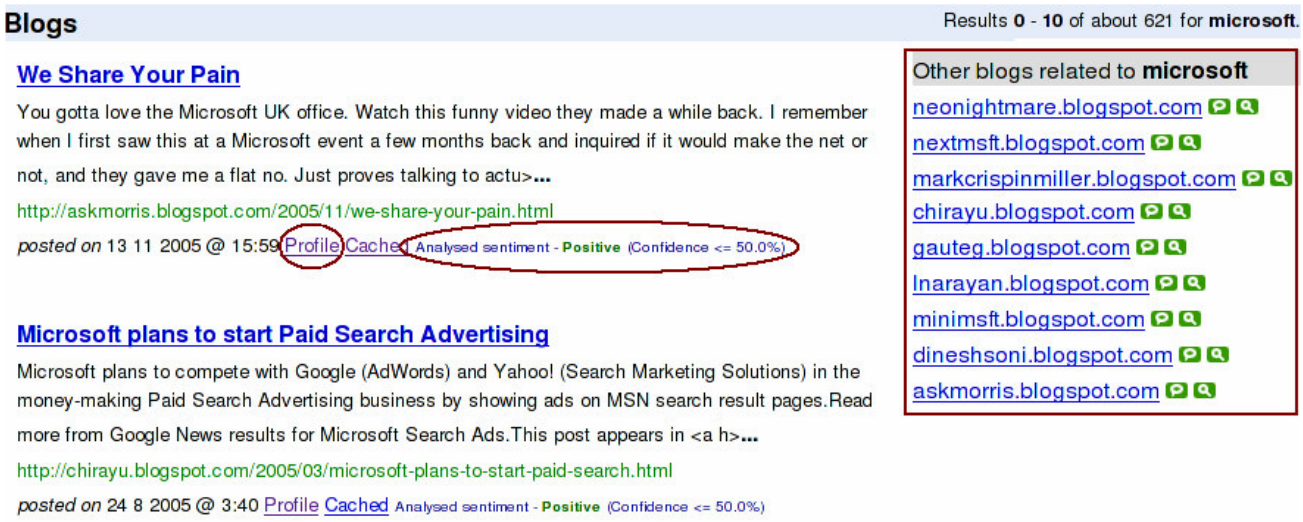

Figure 2 : Search Interface

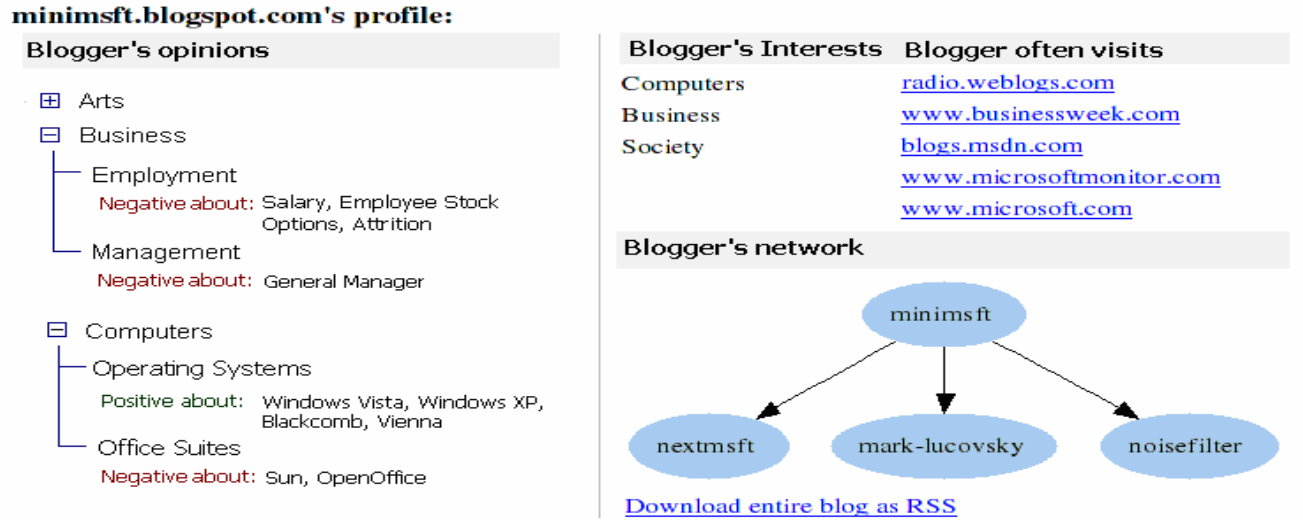

Figure 3 : Blogger's profile

References

[1] Apache Lucene *http://lucene.apache.org*

[2] DMOZ – Open directory project *http://dmoz.org/rdf/*

[3] Herring et. al. Conversations in the blogosphere: An analysis "from the bottom up." Proceedings of the Thirty-eighth Hawaii International Conference on System Sciences (HICSS-38). Los Alamitos: IEEE Press

[4] Google Blog Search *http://blogsearch.google.com*

[5] IMDB: The Internet Movie Database *http://www.imdb.com*

[6] K-nearest neighbour clustering algorithm Ethem Alpaydin – Introduction to machine learning, MIT Press

[7] Marlow C. (2004) Audience, structure and authority in the weblog community. International Communications Association Conference, May 27-June 1, New Orleans LA

[8] Minqing Hu and Bing Liu : Mining and Summarizing Customer Reviews, 2004.

[9] QTag – Part-of-speech tagger for English language *http://www.english.bham.ac.uk/staff/omason/software/qtag.html*

[10] Rainbow statistical classifier. http://www.cs.cmu.edu/~mccallum/bow/rainbow/

[11] Sara Oswly, Sanjay Sood, Kristian J. Hammond: Domain Specific Affective Classification of Documents, 2006.

[12] Technorati Blog search *http://www.technorati.com*

[13] Technorati Weblog: State of the Blogosphere *http://technorati.com/weblog/2006/02/83.html*

[14] Vivisimo : Automatic categorization and meta-search software *http://vivisimo.com/*

[15] WordNet : A lexical database for English language