A Framework for Modeling Influence, Opinions and Structure in Social Media *

Akshay Java
University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250, USA
aks1@cs.umbc.edu

Problem Statement

The Blogosphere provides an interesting opportunity to study social interactions. Blogs provide a channel to express opinions, facts and thoughts. Through these pieces of information, also known as *memes*, bloggers influence each other and engage in conversations that ultimately lead to exchange of ideas and spread of information. We aim to characterize and model the Blogosphere to study the spread of influence (Java *et al.* 2006b), opinion formation and social interaction. Further, we propose a simple generative process that models creation and evolution of blogs. This model is an extension of existing preferential attachment and random surfer model. Using available samples of the Blogosphre (represented as a blog graph) and the proposed generative model, we hypothesize, compare and validate different approaches to modeling influence and opinion formation in social media.

Following are the main contributions of this thesis:

- We propose a novel approach that models influence using topics, polarity of sentiment, bias and temporality.
- We describe a generative model for creating simulated blog graphs with properties similar to the Blogosphere.

Preliminary Work on Opinion Extraction

Opinion extraction and sentiment detection have been previously studied for mining sentiments and reviews in domains such as consumer products (Dave, Lawrence, & Pennock 2003) or movies (Pang, Lee, & Vaithyanathan 2002; Gilad Mishne 2006). More recently, blogs have become a new medium through which users express sentiments. Opinion extraction has thus become important for understanding consumer biases and is being used as a new tool for market intelligence (Glance *et al.* 2005; Nigam & Hurst 2004; Liu, Hu, & Cheng 2005).

The BlogVox system (Java *et al.* 2007a) retrieves opinionated blog posts for topics specified by ad hoc queries. BlogVox was developed for the 2006 TREC blog track and it uses a novel system to recognize legitimate posts and discriminate against spam blogs (Java *et al.* 2006a). Using heuristics and machine learning we processes posts to eliminate extraneous non-content, including blog-rolls, link-rolls, advertisements and sidebars. After retrieving posts relevant to a topic query, the system processes them to produce a set of independent features estimating the likelihood that a post expresses an opinion about the topic. These scores are combined using an SVM-based system and integrated with the relevancy score to rank relevant results. We evaluated BlogVox's performance against human assessors. Our results indicate that data cleaning significantly increased the performance of the system. As shown in Table 1, the overall performance as measured by the *mean average precision* and *R-precision* scores showed that the system worked well on most of the fifty test queries.

Run	Opinion		Topic Relevance	
	MAP	R-prec	MAP	R-prec
Unclean Index	**0.1275**	**0.202**	**0.1928**	**0.2858**
Cleaned Index	**0.1548**	**0.2388**	**0.2268**	**0.3272**

Table 1: The results for the opinion and topic relevance performance of different runs

Some of the challenging questions that are being addressed as part of the ongoing work are as follows:

- Blog posts contain noisy, ungrammatical and poorly structured text. This makes open-domain, opinion retrieval for blogs challenging. We are exploring different techniques for finding topic-specific sentiment words and dealing with slangs and ungramatical text.
- The complexities of human language make it difficult to understand nuances such as contradictions, negations and sarcasm. Such complex linguistic structures require us to rely on semantics rather than shallow NLP. As part of the ongoing work, we have developed Ontological Semantics based tools for processing RSS News feeds[1] (Java, Finin, & Nirenburg 2006). We plan to adapt some of these modules to work with blog data.

*Partial support was provided by NSF awards ITR-IIS-0326460 and TR-IDM-0219649.
Copyright © 2007, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] http://semnews.umbc.edu

Preliminary Work on Influence Models

Studies on influence in social networks and collaboration graphs have explored the task of identifying key individuals who play an important role in propagating information. This is similar to finding authoritative pages on the Web. However, one important difference is that on the Web influence is often a function of topic. For example, Engadget's[2] influence is in the domain of consumer electronics and Daily Kos[3] in politics. A post in the former is less likely to be very effective in influencing opinions on political issues even though Engadget is one of the most popular blogs on the Web.

The other related dimension of influence is readership. With the large number of niches existing on the Blogosphere, a blog that is relatively low ranked can be highly influential in this small community of interest. In addition, influence can be subjective and based on the interest of the users. Thus by analyzing the readership of a blog we can gain some insights into the community likely to be influenced by it. We have implemented a system called Feeds That Matter[4] that aggregates subscription information across thousands of users to automatically categorize blogs into different topics. It also provides a simple scheme to rank blogs using readership-based influence metrics (Java *et al.* 2007b).

An important component in understanding influence is to detect sentiment and opinions. Aggregated opinions over many users is a predictor for an interesting trend in a community. Sufficient adoption of this trend could lead to a 'tipping point' and consequently influencing the rest of the community. We introduce a novel idea of using 'link polarity' using simple sentiment detection techniques. We represent each edge in the influence graph with a vector of topic and corresponding weights indicating either positive or negative sentiment associated with the link for a topic. Thus if a blog A links to a blog B with a negative sentiment for a topic T, influencing B would have little effect on A. Opinions are also manifested as biases. A community of ipod fanatics for example, needs little or no convincing about the product. Thus influencing an opinion leader in such already positively biased communities is going to have less significant impact for the product. Using link polarity and trust propagation we demonstrate how like-mided blogs can be discovered and the potential of using this technique for more generic problems such as detecting trustworthy nodes in web graphs (Kale *et al.* 2007).

Ongoing and Future Work

The first goal of this thesis is to come up with a generic influence model framework. Such a model would incorporate various aspects of influence that were identified as part of the preliminary research. We start by extending existing epidemic based influence propagation techniques (Kempe, Kleinberg, & Tardos 2003) to predict key influencers in a social network. We enhance the influence model to include components for community structure, topic categorization, representation of key beliefs and opinions, and a temporal analysis of how these change. We show that this model outperforms existing, simpler models by evaluating its precision and recall for appropriate tasks.

The second goal of this thesis is to describe a generative process that models creation of new blogs and evolution of post network through citations in the blog graph. This model is an extension of existing preferential attachment model and random surfer model. Consider a rather simplistic view that at each timestep a blog (or a blogger) may either be in a read mode or a write mode. If the blog is in write mode it generates a new post. Additionally, the blog can create a link to an existing post in the network with an empirically determined *linking probability*.

In the read mode each blog performs a *preferential random walk*. The walk begins from the blog homepage and the blogger follows one of the *incident* edges with a probability of (1 -β) or randomly teleport to a different blog in the network. Thus the blogger is more likely to visit a popular blog in its neighborhood, but can also select a less popular blog at random. If u corresponds to the homepage of the blogger, the probability that the walk moves to node v is given by,

$$P_{uv} = \begin{cases} \frac{\beta}{d^+(u)} \cdot \frac{k_v}{|V|} + \frac{1-\beta}{|V|}, & \text{if } (u,v) \in E; \\ \frac{1-\beta}{|V|}, & \text{otherwise} \end{cases}$$

where k_v is the indegree of vertex v. Note that this formulation is also similar to the PageRank algorithm. Another interesting aspect of this modified random walk is that unlike the random surfer model, this model has a *recurrent state* which corresponds to the blogger's homepage. We verify the model and compare the graphs with data from available samples of the Blogosphere (Such as WWE, TREC, etc). A simulated blog graph would be useful in studying the effectiveness of various schemes in our influence models.

[2]http://endgadget.com is a popular electronics blog

[3]http://dailykos.com is a left-wing political blog

[4]http://ftm.umbc.edu

References

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW*, 519–528.

Gilad Mishne, N. G. 2006. Predicting movie sales from blogger sentiment. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*.

Glance, N. S.; Hurst, M.; Nigam, K.; Siegler, M.; Stockton, R.; and Tomokiyo, T. 2005. Deriving marketing intelligence from online discussion. In *KDD*, 419–428.

Java, A.; Kolari, P.; Finin, T.; Mayfield, J.; Joshi, A.; and Martineau, J. 2006a. The UMBC/JHU blogvox system. In *Proceedings of the Fifteenth Text Retrieval Conference*.

Java, A.; Kolari, P.; Finin, T.; and Oates, T. 2006b. Modeling the Spread of Influence on the Blogosphere. Technical report, University of Maryland, Baltimore County.

Java, A.; Kolari, P.; Finin, T.; Joshi, A.; Martineau, J.; and Mayfield, J. 2007a. The BlogVox Opinion Retrieval System. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*.

Java, A.; Kolari, P.; Finin, T.; Joshi, A.; and Oates, T. 2007b. Feeds That Matter: A Study of Bloglines Subscriptions. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*. Computer Science and Electrical Engineering. To Appear.

Java, A.; Finin, T.; and Nirenburg, S. 2006. SemNews: A Semantic News Framework. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, 1939–1940. Menlo Park, CA: American Association of Artificial Intelligence. AAAI Student Abstract Program.

Kale, A.; Karandikar, A.; Kolari, P.; Java, A.; Joshi, A.; and Finin, T. 2007. Modeling Trust and Influence in the Blogosphere Using Link Polarity. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*. Short Paper, To Appear.

Kempe, D.; Kleinberg, J. M.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *KDD*, 137–146.

Liu, B.; Hu, M.; and Cheng, J. 2005. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, 342–351. New York, NY, USA: ACM Press.

Nigam, K., and Hurst, M. 2004. Towards a robust metric of opinion. In *Exploring Attitude and Affect in Text: Theories and Applications, AAAI-EAAT 2004*.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP 2002*.