Web (2.0) Mining: Analyzing Social Media

Anupam Joshi, Tim Finin, Akshay Java, Anubhav Kale, and Pranam Kolari
University of Maryland, Baltimore County, Baltimore MD 21250

Abstract

Social media systems such as blogs, photo and link sharing sites, wikis and on-line forums are estimated to produce up to one third of new Web content. One thing that sets these "Web 2.0" sites apart from traditional Web pages and resources is that they are intertwined with other forms of networked data. Their standard hyperlinks are enriched by social networks, comments, trackbacks, advertisements, tags, RDF data and metadata. We describe recent work on building systems that analyse these emerging social media systems to recognize spam blogs, find opinions on topics, identify communities of interest, derive trust relationships, and detect influential bloggers.

1 Introduction

The past few years have seen the advent of Web-based social media systems such as blogs, wikis, media-sharing sites and message forums. Such Web2.0 systems have a significant amount of user generated content, and have become an important new way to publish information, engage in discussions and form communities on the Internet. Their reach and impact is significant with tens of millions of people providing content on a regular basis around the world. Governments, corporations, traditional media companies and NGOs are working to understand how to adapt to them and use them effectively. Citizens, both young and old, are also discovering how social media technology can improve their lives and give them more voice in the world. We must better understand the information ecology of these new publication methods in order to make them and the information they provide more useful, trustworthy and reliable. Doing this requires mining such systems to find communities based on a combination of topic, bias, and underlying beliefs; authors and blogs are most influential within a given community; from where do particular beliefs or ideas originate and how do they spread; what are the most trustworthy sources of information about a particular topic; and what opinions and beliefs characterize a community and how do these opinions change.

The Blogosphere is part of the Web and therefore shares most of it's general characteristics. It differs, however, in some ways that impact how we can model it and use the model to help extract information from it. The common model for the Web in general is as a graph of Web pages with undifferentiated links between pages. The Blogosphere has a much richer network structure in that there are more *types* of nodes which have more *types* of relations between them. For example, the people who contribute to blogs and author blog posts form a social network with their peers, which can be induced by the links between blogs. The blogs themselves form a graph, with direct links to other blogs through *blogrolls* and indirect links through their posts. Blog posts are linked to their host blogs and typically to other blog posts and Web resources as port of their content. A typical blog post has a set of comments that link back to people and blogs associated with them. Finally, the blogosphere trackback protocol generates implicit links between blog posts. Still more detail can be added by taking into account post tags and categories, syndication feeds, and semi-structured metadata in the form of XML and RDF content. Finally, the text around the link also provides significant information. We believe that adapting and extending the work done by many subcommunities in the data mining arena can help develop new techniques to analyze social media.

In the rest of this article, we discuss our ongoing research in analyzing the Blogosphere and extracting useful information from it. We begin by describing an overarching task of discovering which blogs and bloggers are most influential within a community or about a topic. Pursuing this task uncovers a number of problems that must be addressed, three of which we describe here. They are, recognizing spam in the form of blogs and comments, and developing more effective techniques to recognize the social structure of blog communities.

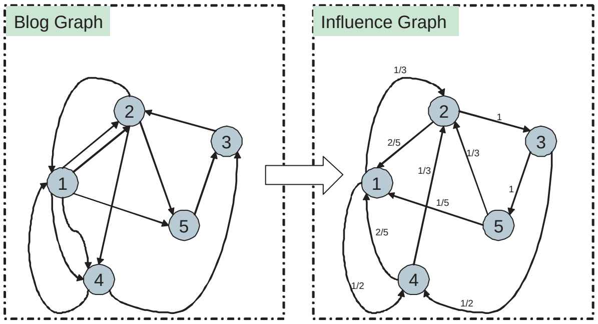

Figure 1. This diagram shows the conversion of a blog graph into an influence graph. A link from *u* to v indicates that *u* is influenced by *v*. The edges in the influence graph are reverse of the blog graph to indicate this influence. Multiple edges indicate stronger influence and are weighed higher

2 Modeling influence in the blogosphere

The blogosphere provides an interesting opportunity to study social interactions including spread of information, opinion formation and influence. Through original content and commentary on topics of current interest, bloggers influence each other and their audience. We aim to study and characterize this social interaction by modeling the blogosphere and providing novel algorithms for analyzing social media content. Figure 1 shows a hypothetical blog graph and its corresponding flow of information in the influence graph.

Studies on influence in social networks and collaboration graphs have typically focused on the task of identifying key individuals who play an important role in propagating information. This is similar to finding authoritative pages on the Web. Epidemic-based models like linear threshold and cascade models [7, 8, 12] have been used to find a small set of individuals who are most influential in social network. However, influence on the Web is often a function of topic. For example, a post contrasting the positions of Repulicans and Democrats on Engadget[1] is not likely to be influential even though it is one of the most popular blogs on the Web. A post comparing the iPhone with the N95 will likely be influential. With Daily Kos[2] the situation will be reversed.

This issue also arises with blogs in niche areas with small readership. A blog that is relatively low ranked on conventional measures 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[3] [4] that aggregates subscription information across thousands of users to automatically categorize blogs into different topics. Table 1 shows the top political blogs ranked using readership-based influence metrics.

1	http://www.talkingpointsmemo.com
2	http://www.dailykos.com
3	http://atrios.blogspot.com
4	http://www.washingtonmonthly.com
5	http://www.wonkette.com
6	http://instapundit.com
7	http://www.juancole.com
8	http://powerlineblog.com
9	http://americablog.blogspot.com
10	http://www.crooksandliars.com

Table 1. The top political blogs ranked using readership-based influence metrics.

An important component in understanding influence is to detect sentiment and opinions. An aggregated opinion over many users is a predictor for an interesting trend in a community, the emergence of a meme. Sufficient adoption of this meme could lead to a 'tipping point' and consequently influence the rest of the community.

Since blog posts are often informally written, poorly structured, rife with spelling and grammatical errors, and feature non-traditional content they are difficult to process with standard language analysis tools. Performing linguistic analysis on blogs is plagued by two additional problems: (i) the presence of spam blogs and spam comments and (ii) extraneous non-content including blog-rolls, link-rolls, advertisements and sidebars. In the next section we describe techniques designed to eliminate spam content from a blog index. This is a vital task before any useful analytics can be supported on social media content.

In the following sections we also discuss 'link polarity' in blog graphs. 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

[1] http://www.engadget.com/

[2] http://dailykos.com/

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

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-minded blogs can be discovered and the potential of using this technique for more generic problems such as detecting trustworthy nodes in web graphs [5].

Existing models of influence have considered a static view of the network. The blogosphere on the other hand is extremely buzzy and dynamic. New topics emerge and blogs constantly rise and fall in popularity. By considering influence as a temporal phenomenon, we can find key individuals that are early adopters or buzz generators for a topic. We propose an abstract model of the blogosphere that provides a systematic approach to modeling the evolution of the link structure and communities. Thus in order to model influence on the blogosphere, we need to consider topic, readership, community structure, sentiment and time.

In the following sections, we provide a brief description of various issues that need to be handled in order to model influence.

3 Detecting blog spam

As with other forms of communication, spam has become a serious problem in blogs and social media, to users and to systems that harvest, index and analyze generated content. Two forms of spam are common in blogs. Spam Blogs, or splogs where the entire blog and hosted posts are machine generated, and spam comments where authentic posts feature machine generated comments. Though splogs continue to be a problem for web search engines and are considered a special case of web spam, they present a new set of challenges for blog analytics. Given the context of this paper and the intricacies of indexing blogs [13] we limit our discussion to that of splogs.

Blog search engines index new blog posts by processing pings from update ping servers, intermediary systems that aggregate notifications from updated blogs. Pings from spam pages increase computational requirements, corrupt results, and eventually reduce user satisfaction. We estimate that more than 50% of all pings are from spam sources [10].

3.1 Detecting Splogs

Over the past year we have developed techniques to detect spam blogs. We discuss highlights of our effort based on splog detection using blog home-pages with local and relational features. Interested readers are referred to [11] [9] for further details.

Results reported in the rest of this section are based on a seed data set of 700 positive (splogs) and 700 negative (authentic blog) labeled examples containing the entire HTML content of each blog home-page. All of the models are based on SVMs [2]. We use linear kernel with top features chosen using mutual information, and models evaluated using one-fold cross validation. We view detection techniques as local and relational, based on feature types used.

3.2 Local Features

A blog's local features can be quite effective for splog detection. A *local feature* is one that is completely determined by the contents of a single web page. A local model built using only these features can provide a quick assessment of the spaminess of blogs. These features include bag-of-words, word n-grams, anchor text, and URLs, We have experimented with many such models, and our results are summarized in Figure 2.

(i) Words. To verify their utility, we created bag-of-words for the samples based on their textual content. We also analyzed discriminating features by ordering features based on weights assigned to them by the linear kernel. It turns out that the model was built around features which the human eye would have typically overlooked. Blogs often contain content that expresses personal opinions, so words like "I", "We", "my", "what" appear commonly on authentic blog posts. To this effect, the bag-of-words model is built on an interesting "blog content genre". In general, such a content genre is not seen on the Web, which partly explains why spam detection using local textual content is less effective there.

(ii) Word N-Grams. An alternative methodology to using textual content for classification is the bag-of-word-N-Grams, where N adjacent words are used as a feature. We evaluated both bag-of-word-2-Grams and bag-of-word-3-Grams, which turned out to be almost as effective as bag-of-words. Interesting discriminative features were observed in this experiment. For instance, text like "comments-off" (comments are usually turned-off in splogs), "new-york" (a high paying advertising term), "in-uncategorized" (spammers do not bother to specify categories for blog posts) are features common to splogs, whereas text like "2-comments", "1-comment", "i-have", "to-my" were some features common to authentic blogs. Similar features ranked highly in the 3-word gram model.

(iii) Tokenized Anchors. Anchor text is the text that appears in an HTML link (i.e., between the `<a...>` and `</a>` tags.) and is a common link-spamming technique around profitable contexts. We used a bag-of-anchors feature, where anchor text on a page, with multiple word

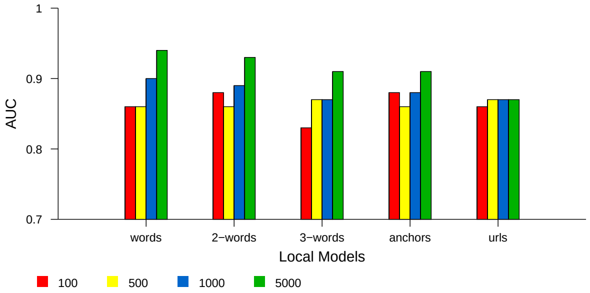

Figure 2. The performance of local models, as measured by the standard, *area under the curve* **metric, varies for different feature types and sizes.**

anchors split into individual words, is used. Note that anchor text is frequently used for web page classification, but typically to classifying the target page rather than the one hosting the link. We observed that "comment" and "flickr" were among the highly ranked features for authentic blogs.

(iv) Tokenized URLs. Intuitively, both local and outgoing URLs can be used as effective attributes for splog detection. This is motivated by the fact that many URL tokens in splogs are in profitable contexts. We term these features as bag-of-urls, arrived at by tokenizing URLs using "/" and ".". Results indicate this can be a useful approach complementing other techniques.

3.3 Relational Features

A global model is one that uses some non-local features, i.e., features requiring data beyond the content of Web page under test. We have investigated the use of link distributions to see if splogs can be identified once they place themselves on the blog (web) hyper-link graph. The intuition is that that authentic blogs are very unlikely to link to splogs and that splogs frequently do link to other splogs. We have evaluated this approach by extending our seed dataset with labeled in-links and out-links, to achieve AUC values of close to 0.85. Interested readers are referred to [11] for further details.

Though current techniques work well, the problem of spam detection is an adversarial challenge. In our continuing efforts we are working towards better addressing concept drift and leveraging community and relational features. The problem of spam in social media is now extending well beyond blogs and is quite common in popular social tools like Myspace and Facebook. The nature of these social tools demand additional emphasis on relational techniques, a direction we are exploring as well.

4 Communities and Polarization of Opinion

Communities, especially on deeply held issues, tend to be likeminded, and often critical of those outside. Our approach uses this idea by associating sentiments with the links connecting two blogs. (By "link" we mean the url that blogger *a* uses in his blog post to refer to blogger *b*'s post). We call this sentiment as *link polarity* and the sign and magnitude of this value is based on the sentiment of text surrounding the link. This can be obtained using shallow NLP techniques. These polar edges indicate the bias/trust/distrust between respective blogs. We then use trust propagation models (in particular, the one by Guha et al.[3])to "spread" the polarity values from a subset of nodes that refer to one another to all possible pairs of nodes. We evaluate the idea of using trust propagation on polar links in the domain of political blogosphere by predicting the "like-mindedness" of democrat and republican blogs. In order to determine if a given blog foo is left or right leaning , we compute the trust/distrust score for foo from a seed set of influential blogs and use a hand-labeled dataset[1] to validate our results. More generally, we address the problem of detecting all such nodes that a given node would trust even if it is not directly connected to them.

We choose political blogs as our test domain; one of the major goals of the experiments was to validate that our proposed approach can correctly classify the blogs into two sets: republican and democrat.

Adamic and Glance [1] provided us with a reference dataset of 1490 blogs with a label of *democrat* and *republican* for each blog. Their data on political leaning is based on analysis of blog directories and manual labeling and has a timeframe of 2004 presidential elections.

Our test dataset from Buzzmetrics[14] did not provide a classified set of political blogs. Hence, for our experiments we used a snapshot of Buzzmetrics that had a complete overlap with our reference dataset to validate the classification results. The snapshot contained 297 blogs, 1309 blog-blog links and 7052 post-post links. The reference dataset labeled 132 blogs as republicans and 165 blogs as democrats (there did not exist any *neutral* labels).

The results in Fig 3 show one of our experiments. They indicate a clear improvement on classifying republican and democrat blogs by applying polar weights to links followed by trust propagation compared to using the link structure alone. Further results and details are described in [5].

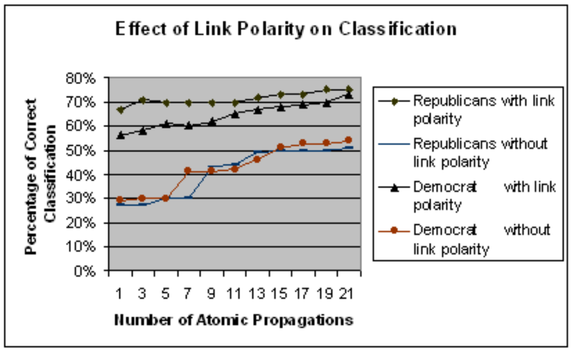

Figure 3. Our experiments show that using polar links for classification yields better results than plain link structure.

5 Conclusion

Social media systems are increasingly import on the Web today and account for a significant fraction of new content. The various kinds of social media are alike in that they all have rich underlying network structures that provide metadata and context that can help when extracting information from their content. We have described some initial results from ongoing work that is focused on extracting, and exploiting this structural information. We note that there is a lack of adequate data sets to fully test the new approaches. To some extent, synthetic blog graphs can be useful to this end. In recent work[6] we have shown that existing graph generation techniques do not work well for this, and proposed a new model.

As the Web continues to evolve, we expect that the ways people interact with it, as content consumers as well as content providers, will also change. The result, however, will continue to represent an interesting and extravagant mixture of underlying networks – networks of individuals, groups, documents, opinions, beliefs, advertisements, and scams. These interwoven networks present new opportunities and challenges for extracting information and knowledge from them.

6 Acknowledgements

We thank Dr. James Mayfield, Justin Martineau and Sandeep Balijepalli for their contributions to the work on sentiment detection, and Amit Karandikar for his work on the generative model. Partial support for this research was provided by the National Science Foundation (awards ITR-IIS-0325172 and NSF-ITR-IDM-0219649) and I.B.M.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, New York, NY, USA, 2005. ACM Press.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, New York, 1992. ACM Press.

[3] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 403–412, New York, NY, USA, 2004. ACM Press.

[4] A. Java, P. Kolari, T. Finin, A. Joshi, and T. Oates. 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, University of Maryland, Baltimore County, March 2007. To Appear.

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

[6] A. Karandikar. Generative Model To Construct Blog and Post Networks In Blogosphere. Master's thesis, May 2007.

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

[8] D. Kempe, J. M. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.

[9] P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[10] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. In *WWW 2006, 3rd Annual Workshop on the Webloggging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[11] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting Spam blogs: A machine learning approach. 2006. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006).

[12] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SIAM International Conference on Data Mining (SDM 2007)*, 2007.

[13] G. Mishne. Applied text analytics for blogs. *Ph.D. Thesis*, January 2007.

[14] NielsenBuzzmetric. http://www.nielsenbuzzmetrics.com.