Modeling Trust and Influence in the Blogosphere Using Link Polarity[1]

Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java, Tim Finin, Anupam Joshi
University of Maryland, Baltimore County
Baltimore MD 21250, USA
{akale1, amitk1, kolari1, aks1, finin, joshi}@cs.umbc.edu

Abstract

There is a growing interest in social network analysis to explore how communities and individuals spread influence. We describe techniques to find "like minded" blogs based on blog-to-blog link sentiment for a particular domain. Using simple sentiment detection techniques, we identify the polarity (positive, negative or neutral) of the text surrounding links that point from one blog post to another. We use trust propagation models to spread this sentiment from a subset of connected blogs to other blogs and deduce like-minded blogs in the blog graph. Our techniques demonstrate the potential of using polar links for more generic problems such as detecting trustworthy nodes in web graphs.

Keywords

Blog, sentiment detection, link polarity, trust propagation

1. Introduction

Social media is a dynamic and growing area that includes collection of blogs, wikis, forums, photos and videos sharing sites. This leads to formation of communities around topics like politics, technology, arts, knitting, photography and public relations. *Influential* nodes in a social network can be responsible for starting a buzz or getting the community to notice a new trend or product. Blogs have become a means by which new ideas and information spreads rapidly on the web. Monitoring and tracking both influential nodes and their opinions on the blogosphere, can thus have a significant number of applications in the realm of product marketing.

In this paper, we address the problem of modeling trust and influence in the blogosphere. Our approach uses links in the blog graph to associate 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. These polar edges indicate the *bias/trust/distrust* between the respective blogs. In order to associate a given blog *foo* to the community of its like-minded blogs, we *create* new *polar links* between all pairs of blogs using initial *polar links*. We use trust propagation models to "spread" the initial polarity values to all possible pairs of nodes. Finally, we compute the trust/distrust score for *foo* from the seed set of *influential* blogs (discussed later) to determine its community. 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.

There has been considerable amount of work in cluster formation and community detection on web graphs, however to our knowledge; none of the prior work involves using polarity of links as a parameter for the problem of community detection. Also, most of the well-known clustering algorithms like [1] are based on the analysis of link structure and may not work well for sparsely connected graphs. Our work is an initial step to address this problem. The remainder of the paper proceeds as follows. Section 2 covers related work. Section 3 describes the details of our approach, heuristic and data modeling. Section 4 covers the experiments and we discuss conclusions and future work in section 5 and 6.

2. Related work

Adar et al. [2] have proposed the use of URL citations to infer the dynamics of *information epidemics* in the blog-space. They also show that the PageRank algorithm finds authoritative blogs. Arun Qamra et al [3] developed a model that incorporates the content of blog entries, their time-stamps, and the community structure to extract the temporal discussions occurring within blogger communities. Ravi et al. [4] have analyzed the word burst models [5] and community structure on the blogosphere [6] and they found a rapid increase in the size of connected component on the blogosphere. Their results on the size of strongly connected components aid in our hypothesis that sentiment detection using text surrounding the links rather than analyzing complete post text has potential for results with high precision-recall. Massa and Avesani [7] have analyzed trust statements in the context of controversial users on social networks.

A number of researchers have worked on the problem of propagating trust in a networked environment. Yu and Singh [8] propose a framework based on the assumptions of symmetric and transitive trust. Kamvar, Schlosser et al [9] have proposed a framework to assign a universal trust to a given node in the graph.. Guha et al [10] in their paper titled *Propagation of trust and distrust* cover work related to trust propagation in multiple disciplines and claim that their work appears to be first "to incorporate distrust in a computational trust propagation setting". We found that their work was most complete and the trust propagation model suits well to our domain.

[1] Partial support was provided by NSF awards ITR-IIS-0326460 and ITR-IDM-0219649 and funding from I.B.M.

To the best of our knowledge, no prior work exists in the area of blogosphere to assign sentiments to links and use such polar links to find like-minded blogs in graph.

3. Proposed approach

In this section, we describe our proposed approach and set the basis for experimental validations. We also provide some details on Guha's trust propagation technique wherever appropriate.

3.1 Link polarity

The term *Link Polarity* represents the opinion of the source blog about the destination blog. VoteLinks[11] is a low case semantic web effort to encode this information directly in html. The sign of polarity (positive, negative or zero) represents whether the bias is for, against or neutral and the magnitude represents how strong or weak the bias is. In order to detect the sentiment based on links, we analyze section of text around the link in the source blog post to determine the sentiment of source blogger about the destination blogger. We consider a window of *x* characters (*x* is variable parameter for our experimental validations) before and after the link.

3.2 Sentiment detection

There has been considerable work on sentiment detection on freeform text. As the first level of approximation, we have not employed any complex natural language processing techniques since bloggers typically convey their *bias* about the post/blog pointed by the link using fairly standard vocabulary. Hence, we use a corpus of positive and negative oriented words and match the token words from the set of *2x* characters against this corpus to determine the polarity.

Since our corpus includes words in noun forms, it is essential for us to employ stemming on tokens. We apply stemming mechanism on all such tokens and then convert them into canonical form by eliminating characters such as commas, periods, exclamation marks etc. Our corpus also includes basic bi-grams of the form "not <positive/negative word>".

3.2.1 Calculation of link polarity

We adopted the following formula for calculating the link polarity:

Polarity = (Np – Nn) / (Np + Nn)

Np : Number of positively oriented words

Nn : Number of negatively oriented words

Notice that our formula incorporates zero polarity links automatically.

3.3 Trust propagation

Since blog graphs may not always be densely connected, we still do not have the trust scores between any given pair of nodes. Hence, we must employ some *sentiment spread* mechanism to calculate trust score between all pairs of nodes from the set of nodes having polar edges between them. Guha et al [10] have proposed a framework to spread trust in a network bootstrapped by a known set of trusted nodes. They have evaluated their approach on a large dataset from epinions[1]. Guha's approach uses a "belief matrix" to represent the initial set of beliefs in the graph. This matrix is generated through a combination of known trust and distrust among a subset of nodes. This matrix is then iteratively modified by using "atomic propagations". Finally "rounding" technique is applied on the matrix thus generated so far, to produce absolute values of trust (yes or no) between all pair of nodes. The "atomic propagation" step incorporates direct propagation, co-citation, transpose trust and trust coupling. We adapt this approach with some modifications for our work. The section on experiments covers our modifications in greater details.

In order to form clusters after the step of trust propagation, we take the approach of averaging trust score for all blog nodes from a predefined set of "trusted" nodes belonging to each community. A positive trust score indicates that the blog node belongs to the community *influenced* by the trusted node of that community. Specifically, we selected top three *influential* democratic and republican bloggers. (We address our notion of *influential* blogs shortly). A positive trust score for a blog *foo* from top three democratic blogs indicates that *foo* belongs to the democratic cluster and a negative score indicates that *foo* is a republic blogger. In order to determine the *influential* bloggers in each community we experimented with the heuristics of high incoming-degree, high outgoing degree and random subset of all nodes.

4. Experiments

We now present the results of our experiments that demonstrate the feasibility and effectiveness of link polarity. Also, we describe the motivation behind choosing the political domain for our experiments and present a representative set of link polarity computations for some of the *influential* blogs.

4.1 Choice of domain

We decided to choose political blogs as our 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 democratic.

Through some manual analysis of the political blogs, we observed that the link density among political blogs is reasonably high and hence we could deduce the effectiveness of our approach by running our algorithms over fairly small number of blogs. In other words, we do not need to perform a large number of iterations of Guha's atomic propagations; about 20 iterations suffice to *create* polar links with sufficiently accurate polarity values between blogs that did not link to each other.

The dataset from Buzzmetrics [12] provides link structure between blog posts over 1.3 million blog posts. Hence, we needed to aggregate this post-post link structure to a blog-blog link structure. This implied that we should choose such a domain where there would be minimal number of off-the-topic posts from the same blog and political blogs fit this requirement perfectly. (We address this issue of determining link polarity based on specific topics in our discussion section).

4.2 Parameters for trust propagation

Guha's work argues that "one step distrust" provides the best trust propagation results in their domain of experiments. They propose the notion of "trust and distrust" between two nodes in the graph where the same set of two nodes can trust or distrust each other. "one step distrust" uses "trust matrix" as the belief matrix. However, we believe that in our domain the initial belief matrix should incorporate both trust and distrust (positive and negative polarities from blog A to blog B). Hence, we use the difference

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

between trust and distrust matrices as our initial belief matrix.

We experimented with various values of the “alpha vector” to confirm that Guha’s conclusion of using the values they proposed {0.4, 0.4, 0.1, 0.1} yields best results. Further, Guha et al recommend performing “atomic propagations” approximately 20 times to get best results; we took the approach of iteratively applying atomic propagations till convergence. Our experiments indeed indicate a value close to 20, after which the final trust scores do not seem to improve. Finally, we do not incorporate the extra step of “rounding” in Guha’s work since the sign of trust is sufficient to determine if the blog under consideration belongs to democratic or republican set.

4.3 Datasets

We studied the effectiveness of our approach over a graph of 300 blogs created from the link structure of buzzmetrics [12] dataset. We observed that in-degree as a heuristic works better over out-degree and random heuristics for selection of *influential* nodes for the seed set. Hence all the results that follow are based on the in-degree heuristic. Lada A. Adamic provided us with a reference dataset of 1490 blogs with a label of democratic or republican for each blog. Some blogs were labeled manually, based on incoming and outgoing links and posts around the time of the 2004 presidential election. Buzzmetrics does not provide a classified set of political blogs. Hence, for our experiments we used a snapshot of Buzzmetrics that had a complete overlap with this reference dataset to validate the classification results obtained by our approach.

4.4 Effect of link polarity

The results in Figure 1 indicate a clear improvement on classifying republican and democratic blogs by applying polar weights to links followed by trust propagation. We get a “cold-start” for democratic blogs and we observe that the overall results are better for republican blogs than democratic blogs. The results being better for republican blogs can be attributed to the observations from [13] that republican blogs typically have a higher connectivity than democratic blogs in the political blogosphere.

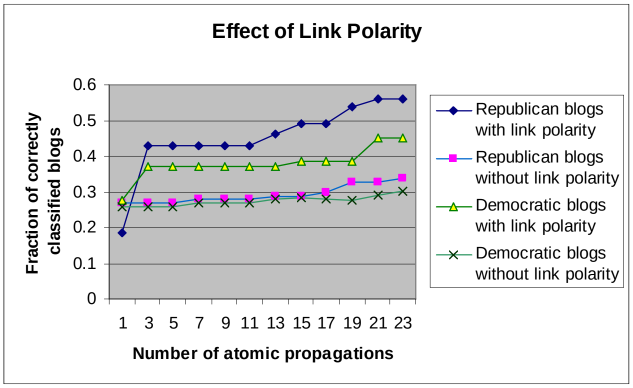

Fig. 1: *Using polar links for classification yields better results than plain link structure*

We are aware of the fact that the results need to be improved further, however it is interesting to note that there exists an upward swing in the accuracy using polar links. Thus, our idea of using trust propagation to *create* polar links between blogs that do not link to each other directly, helps to classify them. This clearly demonstrates the potential of our approach. The linear curve should not be generalized as a typical characteristic of blogosphere, it might be due to certain attributes of our dataset.

4.5 Sample polarity computations

The table in figure 2 depicts polarity values computed between some pairs of *influential* democratic and republican blogs. We present this data as a quick measure of demonstrating the potential of our work and make the following observations.

1. Trust propagation was effective in predicting the accurate polarity for DK-AT, even though our text processing did not yield the correct polarity initially.
2. Trust propagation retained the sign of polarity if the initial computed sign of polarity was correct (e.g., AT-DK). In fact, trust propagation helped in assigning correct polarities to non-existent links (e.g., AT-IP).
3. The numbers in italics indicate the instances where trust propagation failed to assign correct sign to the polarity. However, notice that none of these had any polarity value to start with, so even if trust propagation did not assign the right sign to the link; it helped the clustering process for other blogs by establishing a connection between these blogs. We plan to work on a detailed analysis of such failures in order to get an insight into the effectiveness of our heuristics for link polarity determination. A preliminary analysis indicates that such failures are most likely due to the fact that there are fewer than three links between most blogs in our dataset, hence averaging over such small dataset leads to incorrect sentiment prediction occasionally.

From–To	Num links	Polarity before trust propagation	Polarity after trust propagation
MM-MM	0	N/A	+1.007
MM –DK	0	N/A	-9.290
MM–IP	10	+1.000	+1.370
MM–AT	0	N/A	***+3.530***
DK–MM	0	N/A	-9.290
DK–DK	0	N/A	+8.570
DK–IP	0	N/A	***+9.570***
DK–AT	20	-0.084	+3.260
IP–MM	8	+1.000	+1.030
IP–DK	6	+1.000	***+9.570***
IP–IP	0	N/A	+1.060
IP–AT	0	N/A	-3.640
AT–MM	0	N/A	***+3.530***
AT–DK	5	0.342	+3.260
AT–IP	0	N/A	-3.640
AT–AT	0	N/A	+1.241

MM -http://michellemalkin.com, DK-http://dailykos.com

IP-http://instapundit.com, AT-http://atrios.blogspot.com

Fig. 2: *Polarity values for some influential blogs in our dataset*

4. We realized the need to enforce a lower bound on the number of sentiment words found in our text analysis before performing link polarity computation. Guha's model could have worked better if we had set the polarity to zero for all such cases where *Np* + *Nn* was below two.
5. Our validation techniques did not involve computing trust score for a blog *foo* from *influential* blogs in both communities. This implies that polar links help us by providing multiple ways to find like-minded blogs for *foo*. Thus, AT – IP polarity can correctly classify AT even if AT – MM polarity is incorrect. However, we are working on finding more sophisticated techniques to perform such validations in graphs having more than two communities and hence, we did not rely on non-scalable methodologies for our validations.

5. Discussion

We are aware that we need to analyze results for our approach on a larger dataset. We are also investigating better techniques of validating our results and exploring various heuristics to determine topic of the link. Thus, topic as an extra attribute to the link would give us a fine-grained detail on positive or negative sentiment about a topic over a link and we believe that there are interesting applications of what we would like to term as "topical link polarity". While we are optimistic about our approach, we would like to note that the traditional clustering techniques [1, 14, 15, 16] should be preferred over our approach when the graph is strongly connected. As explained before, the key contribution of our approach lies in classifying the *marginal* nodes (which either do not link or link very sparingly to the tightly connected cluster nodes).

6. Conclusion

We describe a novel approach for classifying blogs into predefined sets by applying positive or negative weights to links connecting the blogs. We validated our approach against a labeled dataset and the preliminary results are promising. We use shallow natural language processing for the text around the links to determine the sentiments of one blog about another. This simple way of sentiment detection augmented by propagating trust using well-known trust models classifies the blogs with decent accuracy. The results demonstrate the potential of using polar links for trust determination problems on web graphs and our future work will be focused on addressing this problem.

Acknowledgements

We thank Nielson BuzzMetrics for making the blog dataset available. We thank Lada A. Adamic and Natalie Glance for allowing us to use their labeled dataset of democratic and republican blogs for our experimental validations.

References

[1] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69:066133, 2004.

[2] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In Workshop on the Weblogging Ecosystem, New York, NY, USA, May 2004.

[3] Arun Qamra and Belle Tseng and Edward Y. Chang. "Mining blog stories using community-based and temporal clustering", CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management, pages 58--67, 2006

[4] J. M. Kleinberg. Bursty and hierarchical structure in streams. Data Min. Knowl. Discov., 7(4):373-397,2003.

[5] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. Computer Networks, 31(11-16):1481-1493, 1999.

[6] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In WWW, 2006.

[7] Paolo Massa and Paolo Avesani. Controversial users demand local trust metrics: An experimental study on epinions.com community. In Proc. of AAAI-05, pages 121–126, 2005.

[8] B. Yu and M. P. Singh. A social mechanism of reputation management in electronic communities. In Cooperative Information Agents, pages 154–165, 2000.

[9] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in P2P networks. In Proceedings of the 12th International World Wide Web Conference, pages 640–651, 2003.

[10] Guha R, Kumar R, Raghavan P, Tomkins A. Propagation of trust and distrust. In: *Proceedings of the Thirteenth International World Wide Web Conference*, New York, NY, USA, May 2004. ACM Press, 2004.

[11] http://microformats.org/wiki/vote-links

[12] NielsenBuzzmetric, www.nielsenbuzzmetrics.com

[13] Lada A. Adamic and Natalie Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005)

[14] J. M. Kleinberg. Bursty and hierarchical structure in streams. Data Min. Knowl. Discov., 7(4):373-397,2003.

[15] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. Computer Networks, 31(11-16):1481-1493, 1999.

[16] C. H. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In WWW, 2006. Mountain Views. In Proceedings of the 2rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wide Web Conference, May 2005.