Modeling Trust and Influence on Blogosphere using Link Polarity

by
Anubhav Kale

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Master of Science
2007

ABSTRACT

Title of Thesis:

Modeling Trust and Influence on Blogosphere using Link Polarity

Author: Anubhav Kale, Master of Science, 2007

Thesis directed by: Dr. Tim Finin, Professor
Department of Computer Science and
Electrical Engineering

There is a growing interest in exploring the role of social networks for understanding how communities and individuals spread influence. In a densely connected world where much of our communication happens online, social media and networks have a great potential in influencing our thoughts and actions. The key contribution of our work is generation of a fully-connected polar social network graph from the sparsely connected social network graph in the context of blogs, where the vertex represents a blogger and the weight of an edge in the polar network represents the bias/trust/distrust between its connecting vertices (the source and destination bloggers). Our approach uses the link structure of 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 post from blogger *b*). We term this sentiment as *link polarity* and the sign and magnitude of this value is based on the sentiment of text surrounding the link. We then use trust propagation models to spread this sentiment from a subset of connected blogs to other blogs to generate the fully connected polar blog graph. Our simple heuristics for analysis of text surrounding links and generation of missing polar

links (links with positive or negative sentiment) using trust propagation is highly applicable for domains having weak link structure. This work has numerous applications such as finding “like minded” blogs, detecting influential bloggers, locating bloggers with specific biases about a predefined set of topics etc. Our experimental validation on determining “like minded” blogs on the political blogosphere demonstrates the potential of using polar links for more generic problems such as detecting trustworthy nodes in web graphs.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Social media has gained increasing popularity over last few years and it has significant contributions in enriching the end-user experience on web. According to wikipedia [1] "social media describes the online tools and platforms that people use to share opinions, insights, experiences, and perspectives with each other". Social media websites provide an easy way for end users to express opinions across a variety of topics and provide techniques to collaborate and share information with each other. They are sometimes referred to as "Web 2.0" since they demonstrate a key characteristic of the next generation of the web – high level of user participation.

Blogs play a vital role in spreading new ideas and information on the web. They typically contain "user generated content", which represents the opinion of bloggers about topics ranging from politics, technology, art to knitting or government policies and public relations. Bloggers typically link to posts from other bloggers and thus serve as a source of networking, interactions and discussions in the social media. Though blogging started as a means of a sharing personal events with friends, there is a growing trend to blog about issues of social interest. This results in other blogs making positive or negative comments for a post and leads to the formation of *links* between different blogs. As in the case of other

[1] http://en.wikipedia.org/

social networking websites such as orkut[2] and MySpace[3], communities around various topics have emerged in the blogosphere. Having an insight into these communities, the temporal factors that affect the buzz about new topics in communities and the knowledge of *influential* bloggers can have a significant number of applications in a number of different areas such as product marketing.

In this thesis, we address the problem of modeling trust and influence in the blogosphere. Our approach uses the link structure of 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. We then use trust propagation models to "spread" the polarity values from a subset of nodes 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 (discussed later) and use a hand-labeled dataset 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.

Our approach uses simple shallow natural language processing to determine link polarity, yet results indicate that our approach has the potential to aid conventional community detection techniques based on path distance and reachability metrics. Since, we do not process entire blog-post text for sentiment detection between two blogs and use shallow NLP techniques, we speculate that the approach should scale well for real-time applications (e.g., analyzing blogs for dynamic situations like elections) than traditional off-line and

[2]http://www.orkut.com/

[3]http://www.myspace.com/

computationally intensive approaches. This work presents some of our interesting results in the domain of blogosphere, however a long-term goal of our work is to deduce trustworthy nodes for a given node in any web-graph. We believe that directed polar links have a tremendous potential for addressing this hard problem.

1.1 Background and Related Work

Bloggers typically discuss views about varied topics and are based on personal experiences. Such views are expressed almost instantaneously as soon as any new event occurs. The blogosphere has matured a lot since its inception and hence, when an event occurs, the first reaction is to turn to the blogosphere to see what people are saying about it. For example, during the London bombings in 2006 people were interested in finding first hand reports, pictures, emotions and experiences of Londoners. As time progressed, people might have looked for more information about the event - what happened? Why? How many people were killed? Have there been any arrests? Which group(s) has claimed responsibility for this act? etc.

Suppose that your goal was to market a new kind of mp3 player which would compete with ipod. One of the starting points would be to use advertising products such as Google's popular Adsense[4] product. Using content of the webpage, this service matches relevant web pages to advertisements that relate to the topic of the page. While this gives a wide coverage and a significant audience, there is very little the advertiser can do to actively promote the product to the right set of individuals. Using a blog search engine one can find a ranked list of relevant blog posts for different generic query terms. However, most blog search engines use link based ranking schemes that measure popularity as opposed to influence. While a number of popular blogs may talk about ipods in general, if the marketing division of your company can target the community that has a negative bias

[4]https://www.google.com/adsense/

about ipod then chances of spreading good word about the new mp3 player is considerably high than targeting the community having a strong positive bias about ipod already. Thus having an insight into the communities in social media can aid in accurately targeting key personnel for marketing new products.

Temporal analysis of the swing in trends among communities has interesting applications for scenarios such as elections where a study of cause and effect phenomena has tremendous potential to gain an insight into change in voters' (or bloggers') bias during the election campaign events. This further implies that a community detection system capable of performing highly efficient real-time analysis of streaming data from social media can play an important role for analyzing the effects of a candidate's meetings and speeches during the election time.

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 do not work well for sparsely connected graphs. We believe that our contributions can be applied in the domain of community detection as well.

The remainder of the thesis is organized as follows. Chapter 2 covers related work. Chapter 3 describes the details of our approach, heuristic and data modeling. Chapter 4 covers the experiments and we discuss conclusions and future work in Chapter 5.

Chapter 2

RELATED WORK

In this chapter, we cover the research work related to our contributions. Since our work spans across various fields of research in the computer science community, the related work is organized into different sections based on the broad area of research.

2.1 Sentiment Analysis

Sentiment analysis can be defined as determining the overall *polarity* of a given document. The motivation behind this work can be attributed to various factors ranging from individual-centric applications such as determining positive or negative reviews for a product to more commercial business models like recognizing and discarding "flames" on newsgroups, analyzing opinions on government policies etc. Researchers have focused on many interesting challenges in this area such as predicting correct polarity irrespective of references to different objects in the same text corpus, modeling the context of text for topic categorization, analyzing language specific nuances such as negated words, n-grams, metaphors and subtle expressions; to name a few.

Turney [2] propose a simple unsupervised learning algorithm for classifying reviews on the web as "thumbs up" or "thumbs down". Turney's work is focused on using the "semantic orientation" of phrases which is calculated as the difference between the mutual information gain between a given phrase and "excellent" and the mutual information

gain between the same phrase and “poor”. This work provides a simple, yet effective way of handling the complex natural language processing problem of sentiment classification. Pang et al. [3] provide a detailed analysis of various machine learning algorithms for the movie review classification problem. Their analysis of the “thwarted expectations” in the domain of movie reviews indicates yet another challenge in the domain of sentiment classification.

Church et al. [4] present work on “word association norms” (classifying words based on the co-occurrence with other words in corpus). Their approach uses information theoretic models of mutual information for estimating word association norms. Models based on information theory are more effective than the traditional and costly way of testing few thousands of subjects on few hundred words to determine word associations. Though this work is very theoretical, it has many applications in enhancing the productivity of lexicographers. Das and Chen [5] present a manually created lexicon and various scoring techniques to classify postings in stock message boards. Though the classification accuracy of their approach is close to Bayes classifiers, their “noise reduction” techniques reduces false positives to a great extent. We believe that their work is very complete in terms of analysis of classifier algorithms, voting mechanisms and the wide range of metrics. Liu et al. [6] in their work on “Opinion Observer” propose a pattern mining technique to extract features of the product from Pros and Cons of a review. Their work differs from the various models suggested for polarity determination in the granularity of semantic processing, since they can identify the *features* that the customers praise or complain about.

Hearst [7] uses cognitive linguistics to determine the *directionality* of a sentence. This approach is a loose-case effort for applications that do not have sufficient resources to engage into complex NLP techniques, however the approach is useful only if the cost of building and executing the proposed methods does not compromise the quality of results. The work is independent of any domain-specific ontologies and uses isolated text interpre-

tation in the realm of a generic metamorphic model adopted from [8]. Pang and Lee [9] use minimum cuts in graphs to extract the subjective text from the corpus under interest. Use of minimum cuts for polarity classification is a novel technique and helps in efficient text mining.

2.2 Community Detection

Community detection or clustering can be defined as *the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure* [1]. "Distance measure" is a term used to determine how the *similarity* between given elements is calculated for the process of clustering. Thus, community detection belongs to the general class of problems that deal with inferring knowledge from data.

The structure of blog includes blogrolls (links to other blogs that this blogger follows regularly), links to other blog posts and comments from other bloggers which collectively establishes a "web of hypertext" automatically. Gurak et al. [10] provides case study of a blog called "Julie/Julia Project"[2] to determine if a blog in itself can have a sense of "virtual community". They conducted a survey of the frequent readers of the blog and concluded that the blog was more of a "virtual settlement" than a "virtual community". Nevertheless, their work indicated that blogs do have the potential to create communities on the web. Efimova et al. [11] use blog readership patterns, data on blogs linking to each other, *conversations* between bloggers and blog directories as indicators of communities in the blog-space. They prove the existence of "mini-clusters" or marginal nodes in every significant blog community and comment on the reasons behind such marginal cluster formations.

[1] http://en.wikipedia.org/wiki/Data_clustering
[2] http://blogs.salon.com/0001399/

Herring et al. [12] study the extent to which the blogs are interconnected using *A-List Blogs* as the seed set for bootstrapping. They aggregated top 100 blogs from NITLE Blog Census [3], Technorati [4], and the TTLB Blogosphere [5]. They use social network analysis techniques of distance metrics, bi-directional links, in-degree and out-degree to deduce the connectivity and conversation patterns in blogs and conclude that the blogosphere is "partially interconnected and sporadically conversational".

Tyler, Wilkinson and Huberman [13] use e-mail as the domain for community detection. They evaluate an algorithm based on the centrality measure to deduce formal and informal communities over a corpus of one million e-mail messages within the HP Labs. Fisher [14] takes a more individual-centric approach in the form of *egocentric social network analysis* where data pertaining to only the immediate neighbors for any user is analysed. This approach alleviates the typical problems of social network analysis like privacy concerns, not having sufficient privileges to access proprietary data, overheads of extensive data collection and such. They use *out degree histograms* to demonstrate the potential of their "single perspective" approach on the domain of e-mail and newsgroups and comment on its applications in e-mail visualisation.

2.3 Trust Representation and Propagation

People on Web 1.0 and software agents on Web 2.0 have to interact with unknown entities (*strangers*) to accomplish a variety of online tasks. Most of the commercial e-commerce websites in Web 1.0 (e.g. Amazon [6]) rely heavily on models for representing trust based on ranking schemes. Since it is not practical for every entity (people or software agents) on the web to have explicit knowledge of trust about every other entity, it

[3] http://www.blogcensus.net/
[4] http://www.technorati.com/
[5] http://truthlaidbear.com/
[6] http://www.amazon.com

is required to *predict* the trust score for a stranger from the knowledge of trust scores for known (or trusted) entities. Researchers have focused on this problem of formally modeling the notions of trust, distrust, influence and techniques for deriving trust scores for unknown entities. There has been considerable amount of work in disciplines other than computer science on various aspects of trust definitions, trust metrics, trust propagation models and validation techniques.

Huang and Fox [15] provide a formal framework for representing trust and study the transitivity of trust. They classify trust as "trust in belief" and "trust in performance" and prove the transitivity of "trust in belief". They define the concept of trust as *Trust is the psychological state comprising (1) expectancy: the truster expects a specific behavior of the trustee such as providing valid information or effectively performing cooperative actions; (2) belief: the truster believes that expectancy is true, based on evidence of the trustees competence and goodwill; (3) willingness to be vulnerable: the truster is willing to be vulnerable to that belief in a specific context where the information is used or the actions are applied.*. To the best of our knowledge, this is the most precise and complete definition of trust since it provides a domain-independent abstraction for the definition of trust. Gans et al. [16] argue for the importance of giving an explicit consideration for distrust on social networks. They propose a "TCD" model based on the notion of trust, network confidence and distrust. The idea of using "network confidence" as a parameter in social network simulations has interesting applications. Beth, Borcherding and Klein [17] have worked on the problem of determining the trust for an entity in the context of conflicting recommendations about its trustworthiness. They emphasize that the semantics of direct trust values differ from that of the recommended trust values. Their mathematical models for combining conflicting trust scores are based on the non-monotonicy property of trust and use arithmetic mean as the mode of aggregation.

Richardson, Agarwal and Domingos [18] have proposed a framework to represent

trust and distrust on the semantic web. Their idea is to compute a subjective trust/distrust score for every user instead of assigning a global trust score to every user. They use "linear pool", "noisy OR" and "logistic regression" for the combination functions and their work is one of the few which is evaluated on a very large dataset from Epinions[7]. Yu and Singh [19] study the problem of adversaries in trust management systems. They provide a model to detect deception in the process of trust/distrust propagation in a networked environment and we believe that their models are very applicable to social networks. Kamvar, Schlosser and Garcia-Molina [20] propose a secure method to calculate global trust values for shared files in P2P networks. The goal of their "Eigen Trust" algorithm is to reduce downloads of inauthentic files using global trust scores assigned to each file. Guha et al [21] in their work 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. Hence, our trust propagation approach is very similar to their work.

2.4 Miscellaneous Social network Analysis

We believe that the research in the area of information propagation was inspired by a large body of work in disease and epidemic propagation. Gruhl, Guha, Nowel and Tomkins [22] study the features of information propagation through the blogspace. Their study is focused on two key aspects of information spread viz. the *topic* and the *individual*. They formalize the idea of topics that run over long period of time and use theory of infectious diseases to analyze the flow of information. They further classify the long running topics as internal sustained discussion and externally induced spikes and provide formal models for both of them. They propose an "expectation-maximization" algorithm which predicts

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

the probability of an individual getting infected by a topic at a given epoch of time. Adar et al. [23] study macroscopic and microscopic patterns of blog epidemics and propose ranking algorithms that take advantage of infection patterns. They try to find co-relations between different types of information and epidemic profiles and model citation signals as an approximation of information epidemics. They describe a variation of page rank algorithm termed as “iRank” which ranks blogs based on their informativeness. “irank” is based on a function of temporal proximity of URL citations.

Since blogs contribute to the process of spreading buzz about new topics on the social networks, recent research has focused on extracting opinions and communities from blogs [24]. Arun Qamra et al [25] have developed a “Content Community Time” model which uses the blog-post content, its timestamp and community structure of blog-space to infer temporal discussions or *stories* on the blog-space. Gruhl et al. [26] demonstrate that online postings can be used to predict spikes in sales ranks. They create hand-crafted queries from 3,00,000 blog postings and evaluate them on the Amazon[8] sales data . They study various predictor algorithms (weighted average, markov and weighted least-square) for the experimental evaluations and comment about the effectiveness of each. However, we believe that their work is very specific to the Amazon dataset and may not be applied directly to other domains such as electoral voting schemes, public opinion on government policies etc. Lloyd et al. [27] have compared the content of popular blogs and major U.S. newspapers over the same timeframe and derived the correlation between news and blog references. They prove that the “lead/lag” shifts required for comparing the *stories* in both sources vary to a great extent depending upon the type of *story*. Kumar et al. [28] found rapid growth in the size of connected component on the blogosphere. They argue that this trend is due to the increasing tendency of bloggers to comment about other blogs. Their results on the size of strongly connected components aid in our hypothesis that sentiment

[8]www.amazon.com

detection using links rather than analyzing the post text has potential for results with high precision-recall.

Tags in popular Web 2.0 websites provide an effective data-organization technique for end users. In recent years, tags have become very popular and researchers have coined a term *Folksonomy* to represent the tag cloud in social media. Brooks and Montanez [29] have analyzed tags from Technorati[9] to calculate the similarity of all articles having a common tag. They use clustering algorithms to deduce "topical hierarchy" in tags and argue that tags are effective in grouping similar articles but not very effective for determining the content of an article. Shen and Wu [30] model folksonomy as a "tag network" . According to them, the model constructed using the sampled folksonomy from Del.icio.us[10] demonstrates the properties of "small world" and "scale free" network. Marlow et al. [31] study popularity and influence as the measures of authority on the blogosphere. Their analysis indicates that permalinks can serve as a good approximation of influence; however it is important to consider the distribution of permalink citations for each blog in order to normalize the variations of incoming traffic on different posts in the same blog. Present blog search engines do use permalink citations to determine the relevance and authority of search results.

To the best of our knowledge, no prior work exists in the area of blogosphere to assign sentiments to links (what we term as "link polarity") and use trust propagation on such polar links to find "like-minded" blogs.

[9] http://www.technorati.com
[10] http://del.icio.us/

Chapter 3

PROPOSED APPROACH

In this chapter, we describe our proposed approach and set the basis for experimental validations. We also provide 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. 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.

Bloggers can *link* to each other in one or more of the following ways.

1. **Explicits Links**

   Bloggers typically provide links to the other blogs in their blog posts. We believe that such explicit links are the strongest evidence of interaction between bloggers. The very fact that a blogger takes efforts and spends time to provide explicit link to other blog indicates that he is genuinely interested in expressing his opinion about other blog. From observations on our data-set of political blogs, we found that the probability that a blogger expresses some opinion in his blog-post about the *linked* blog is very high and this serves as an effective measure for computing the *polarity* of the association between bloggers.

2. **Comments**

   Most of the blogging softwares provide mechanisms to write comments on blog-posts. However, we observe that it is a common tendency to comment anonymously on controversial topics. Also, comments can be treated as an opinion at a higher granularity than explicit links, since they provide feedback on the entire blog-post as against specific parts of it. If the comment is too verbose and addresses various topics in the blog–post, then it is difficult to use simple NLP techniques to compute its *polarity*. Further, we feel that comments can be treated as "pull model" where the source blogger may not necessarily be associated with the bloggers commenting on it and hence comments do not serve as an acurate measure of trust/distrust between bloggers.

3. **Blogrolls**

   Blogrolls is a direct measure of judging the *association* between bloggers. However, bloggers can include a particular blog in their blogroll for a variety of reasons including friend or family relationships. Thus, the presence of a blog in blogroll does not necessarily indicate positive bias about the blog, it just indicates that the blogger is *interested* in following this blog. Also, bloggers typically do not update the entries in their blogrolls which makes blogroll entries very *static*. However, we do feel that associating a positive value to all blogs in the blogroll may help in our step of trust propagation if there is a contradictory or low evidence of trust/distrust from the link polarity computations.

In order to determine 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. From our analysis of blog texts and interactions with regular bloggers, we observed that it is not necessary to analyze the complete blog post text to determine the sentiment. In fact, text neighboring the link provides direct meaningful insight into blogger

a's opinion about blogger *b*. Hence, we consider a window of *x* characters (*x* is variable parameter for our experimental validations) before and after the link. Note that this set of *2x* characters does not include html tags.

3.2 Sentiment Detection

There has been considerable work on sentiment detection on free–form text. Researchers have experimented with various approaches for text processing on the web and we have covered the relevant research in the chapter on related work. For our requirements, we do not need to employ 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. We observed that bloggers frequently use negation of sentimental words while indicating bias about another blog-post("What b says is not bad"), hence our corpus also includes basic bi-grams of the form "not positive/negative word". Our experiments confirmed that the aforementioned simple techniques are very effective in deducing the text sentiment correctly.

3.2.1 Canonicalization of tokens

Since we use a static list of words having positive and negative orientations, it is important to convert the words in our text window into canonical form to achieve maximum match with the corpus words. We apply stemming – *"process for removing the commoner morphological and inflexional endings from words in English"* on the tokens using the Porter [32] stemming algorithm. The third-party implementation that we use deviates from the standard algorithm. However, it does not affect the recall for our system since we are not concerned with any linguistic exercise. Also, the algorithm automatically handles

special cases such as not removing the suffix for very short stems.

3.2.2 Calculation of link polarity

The number of positive and negative words varies to a great extent (typically from 1 to 30 in window size of 750 characters) across multiple posts. Hence, it is necessary to normalize the results over some metric. We adopted the following formula for calculating the link polarity between two posts:

$$Polarity = (Np - Nn) / (Np + Nn).$$

$$Np : Number\ of\ positively\ oriented\ words$$

$$Nn : Number\ of\ negatively\ oriented\ words$$

Notice that our formula incorporates zero polarity links automatically. The term in the denominator ensures that the polarity is weighed according to the number of words matched against the corpus. This helps to differentiate all such instances where (Np – Nn) is the same but (Np + Nn) varies from a small value (minimum = 2) to a large value (typically, 20). Also, note that we do not incorporate the number of links in our polarity computation.

We use summation as the aggregation technique for computing the polarity between two blogs. For our experiments, we choose a domain with a low probability of "off-the-topic" posts within a single blog, hence the notion of summing post-post polarity values to yield a blog-blog polarity value holds good. We will have to investigate better aggregation techniques for handing more noisy datasets or filter the dataset and then apply the method of summation.

3.3 Trust Propagation

Since blog graphs are not densely connected [12], we still do not have the trust scores between any given pair of nodes. Hence, we must employ some sentiment spread mecha-

Table 3.1. Matrix representations

Name	Representation
T	Trust between users *i* and *j*
D	Distrust between users *i* and *j*
B	Initial Belief between users *i* and *j*
C	Atomic Propagation Operator
F	Final Belief between users *i* and *j*

nism to calculate trust score between all pairs of nodes from the set of nodes having polar edges between them. As explained in the chapter on related work, we use Guha et al's [21] work for trust propagation and present a detailed summary of their work in the following section.

3.3.1 Summary of Guha et al's work

Guha et al [21] 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].

The framework consists of n users and n * n matrices described in Table 4.1 to model trust representation and trust propagation. B is computed based upon T, D or a combination of trust and distrust viz. T – D. This matrix represents belief values for a subset of all users and the goal is to compute the final belief matrix F which represents beliefs between all pairs of users. F is calculated from B by performing multiple steps of trust propagations termed as *atomic propagations*.

As described by Guha et al., "Atomic Propagation extends a conclusion (such as the conclusion that i trusts j) by a constant-length sequence of forward and backward steps in the graph of expressed trusts". An atomic propagation is based on the following models of trust and distrust propagation. (For sake of simplicity, we use the term "trust" to represent

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

positive and negative bias.)

1. **Direct Propagation**

   Hypothesis: *i* trusts *j* and *j* trusts *k*

   Conclusion: *i* trusts *k*

   Mathematical Representation: B

2. **Co–citation**

   Hypothesis: *i* trusts *j* and *k* and *m* trusts *k*

   Conclusion: *m* trusts *j*

   Mathematical Representation: $B^T * B$

3. **Transpose Trust**

   Hypothesis: *i* trusts *j* and *k* trusts *j*

   Conclusion: *k* and *i* trust each other

   Mathematical Representation: B^T

4. **Trust Coupling**

   Hypothesis: *i* trusts *j* and *m* trusts *i* and *k* trust *j*

   Conclusion: *m* trusts *k*

   Mathematical Representation: $B * B^T$

The combined atomic propagator is defined using a weight vector α as follows.

$$C = \alpha_1 B + \alpha_2 B^T B + \alpha_3 B^T + \alpha_4 B B^T$$

Using the atomic propagator matrix, the belief matrix is computed iteratively. Thus the belief matrix in *i+1* th propagation step (B_{i+1}) is computed from the belief matrix in *i* th propagation step as follows.

$$B_{i+1} = B_i * C_i$$

The total number of atomic propagations required for “convergence” depend on the diameter of graph under consideration. By “convergence” we mean a state where the change

in trust scores for successive propagations is below a pre-defined threshold. In the discussion that follows, *k* represents the number of propagations required for convergence.

Guha et al suggest the following models for computing the final belief matrix F.

1. **Trust only**

   In this case, only trust scores are propagated.

   $F = B_k$, $Bo = T$

2. **Propagated Distrust**

   In this case, both trust and distrust are propagated iteratively.

   $F = B_k$, $B_o = T - D$

3. **Single-step Distrust**

   This case argues that distrust propagates a single step. So if a user *i* distrusts *j*, then bias of *j* about any user unknown to *i* should not affect *i*'s bias about the unknown users.

   $F = B_k * (T – D)$, $B_o = T$

Single-step distrust provides best results on Guha's dataset. However, we believe that such results may not be generalized on arbitrary datasets. Hence, we evaluate all the aforementioned models on our dataset and comment about the results in the chapter on experiments. Guha further provide techniques to *round* the trust scores to +1, 0 and -1. We do not employ rounding for our specific experiment of classifying blogs into like-minded sets.

3.4 Classification of "Like-minded" Blogs

In order to determine the "like-minded" blogs 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* democrat and republican bloggers. (We address our notion of *influential* blogs shortly). A positive trust score for a blog *foo* from top three democrat blogs indicates that *foo* belongs to the democrat cluster and a negative score indicates that *foo* is a republic blogger. Notice that negative links thus help us to classify a blog into the right cluster even if it is not very well connected within its cluster. 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.

Chapter 4

EXPERIMENTS

In this chapter, we cover the details of our experiments to 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 influential blogs from our dataset.

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 democrat. Following are the primary reasons for selection of this domain.

- 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 are sufficient to create polar links between blogs that did not link to each other directly.

- We used the dataset from Buzzmetrics [33] which provides link structure between

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. (In the chapter on conclusion, we address this issue of determining link polarity based on specific topics).

- From a business model point of view, political blogs are highly effective during the timeframe of elections to determine the trends among voters and a technique that can classify voters into multiple political biases would be extremely beneficial to various sources.

4.2 Experimental Parameters

4.2.1 Link Polarity

As explained in chapter 3, we used various window sizes around the links to fetch the token words to be used for sentiment detection. After some manual analysis of political blogs, we decided to experiment with 1000, 750, 500, 250 and 50 characters before and after the link under consideration. We expected to get some insights into what would be the correct window size (and hence, an approximation for the number of words around links that yield more signal then noise) by varying this parameter.

4.2.2 Trust Propagation

We now provide details of our modifications/selection of models from Guha's work. 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 between trust and distrust matrices as our initial belief matrix. We believe that the idea of using "eigenvalue propagation" to determine final trust scores is generic and applies to any domain. Hence we used the same for our experiments.

We experimented with various values of the α *vector* to confirm that Guha's conclusion of using the values they proposed $\{0.4, 0.4, 0.1, 0.1\}$ yields best results. Our experiments indeed confirmed that this set of values yields the most accurate results. In order to evaluate the impact of direct trust, co–citation, transpose trust and trust couping, we experimented with all possible permutations of setting each of the parameter in the α vector to its optimal value and 0. We represent the parameters as a bitset containing 4 bits where *1* represents that the optimal value of the parameter was used and *0* represents that the parameter was discarded. The bits from MSB to LSB represent direct trust, co–citation, transpose trust and trust couping respectively. (e.g. The representation "1101" indicates α = $\{0.4, 0.4, 0 , 0.1\}$)

Further, Guha et al recommend performing "atomic propagations" approximately 20 times to get best results. Since, we can not guarantee that such numbers would work in our domain; 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 democrat or republican set.

4.2.3 Influential Node Selection

There can be a variety of heuristics for deducing an influential node (blog) in a given set of like-minded blogs. We decided to choose the most intuitive ones viz. high incoming

links, high outgoing links, random set and experimented with all of them.

4.3 Datasets

4.3.1 Test Dataset

We studied the effectiveness of our approach over a blog graph created from the link structure of buzzmetrics [33] dataset. The dataset consists of about 14M weblog posts from 3M weblogs collected by Nielsen BuzzMetrics for May 2006. The data is annotated with 1.7M blog-blog links [34]. For our experiments, we use the posts in English which form 51% of the overall post-post data. This dataset was released with the ICWSM conference,2007 [35].

4.3.2 Reference Dataset

Lada A. Adamic 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. Some blogs were labeled manually, based on incoming and outgoing links and posts around the time of the 2004 presidential election.

Our test dataset from Buzzmetrics 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).

4.4 Experimental Results

We now present the results of our experiments. The following sections demonstrate how varying various parameters affects the accuracy of classification. While studying the

effect of a particular parameter, we set the values of other variable parameters to their optimal values (which in turn were obtained by the experiments to study the effect of those parameters).

4.4.1 Link Polarity

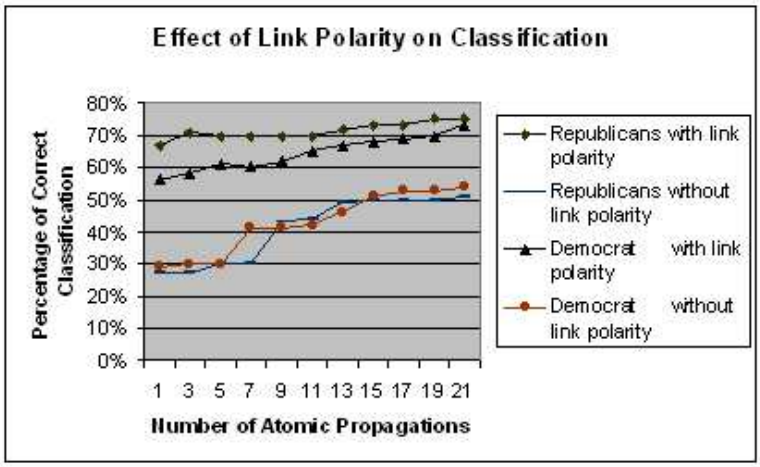

FIG. 4.1. Using polar links for classification yields better results than plain link structure.

The results in 4.1 indicate a clear improvement on classifying republican and democrat blogs by applying polar weights to links followed by trust propagation. We used the heuristic of "'high incoming links"' for influential blog selection and "1111" as the bitset for atomic propagation. We get a "cold-start" for democrat blogs and we observe that the overall results are better for republican blogs than democrat blogs. The results being better for republican blogs can be attributed to the observations from [36] that republican blogs typically have a higher connectivity then democrat blogs in the political blogosphere.

We are aware 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. We

would like to note that the linear curve should not be generalized as a typical characteristic of blogosphere, it might be due to certain attributes of our dataset. We briefly discuss about further analysis of such trends in the chapter on conclusions.

4.4.2 Text-Window Size

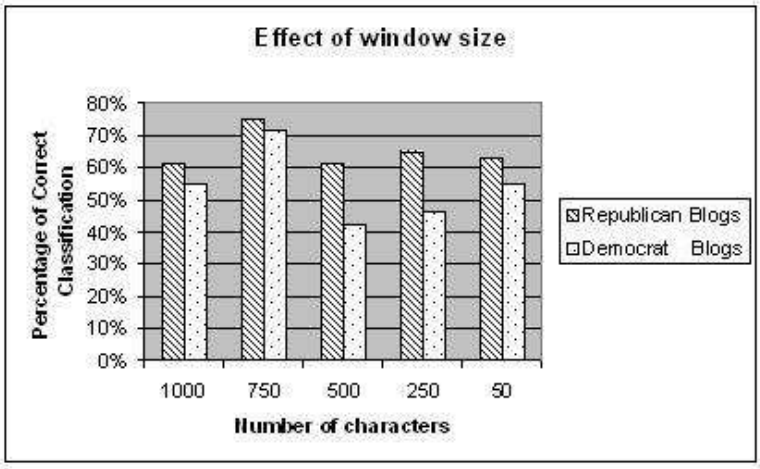

FIG. 4.2. The correctness of classification depends on the optimal window size (around 750 characters) and decays on both sides of the optimal window.

The results in figure 4.2 indicate that 750 characters was the most appropriate window size for our dataset. We used the heuristic of "'high incoming links"' for influential blog selection and "1111" as the bitset for atomic propagation. If the window size is too small, our system becomes susceptible to short non-opinionated phrases around the link (e.g. *here is what xyz says*) which leads to a zero match of token words to corpus words in text surrounding link. On the other hand, if the window size is too large, our system becomes susceptible to analyzing text unrelated to the opinion expressed around the link. Another source of misinterpretation is the presence of other links in our window. Hence we stop extending the window from the link whenever we hit the window size *x* or another link at a distance of atleast 50 characters from the link under consideration. (The heuristic of 50

characters is incorporated to avoid missing text in such cases where bloggers clutter *related* links very close to the link under consideration.)

4.4.3 Evaluation Metrics

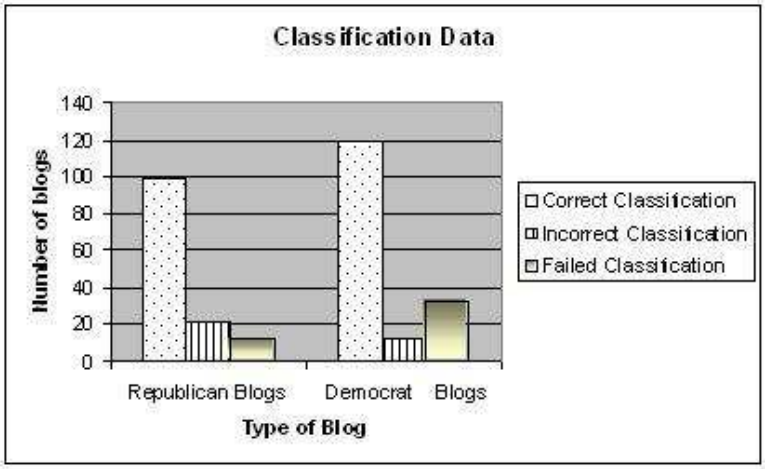

FIG. 4.3. Trust Propagation on polar links yields blog classification with high accuracy

		Predicted	
		Democrat	Republican
Actual	Democrat	99	45
	Republican	33	120

FIG. 4.4. Confusion Matrix

We now discuss various evaluation measures computed from Figure 4.4 and using definitions from [37]. We use *positive* to denote republican blogs and *negative* to denote democrat blogs in order to comply with standard terminology.

Accuracy = 73%

True Positive Rate (TP) = 78%

False Positive Rate (FP) = 31%

True Negative Rate (TN) = 69%

False Negative Rate (FN) = 21%

Precision (Positive) = 75%

Precision (Negative) = 72%

4.4.4 Atomic Propagation Parameters

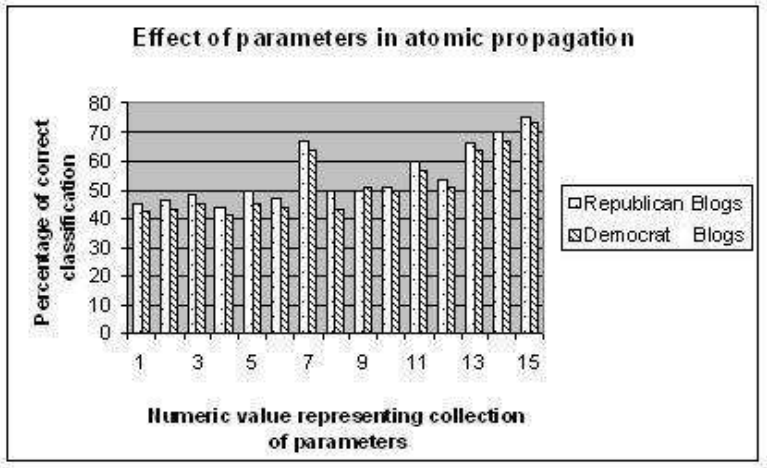

FIG. 4.5. Selection of atomic propagation parameters dominates classification accuracy

The results in figure 4.5 demonstrates how the percentage of correct classification depends on the acurate selection of atomic propagation parameters. Though, it is intuitive that selecting all parameters would yield the best results, it is an informative excercise to vary all parameters and study the results. For this evaluation, we used the heuristic of "High incoming degree" for influential node selection.

4.4.5 Heuristics for Selection of Influential Nodes

Since trust propagation is inherently a "push" model in which the trust/distrust is pushed from a subset of nodes to all nodes, high out-going degree seems to be the best heuristic for influential node selection. However, as the results in figure 4.6 indicate, high

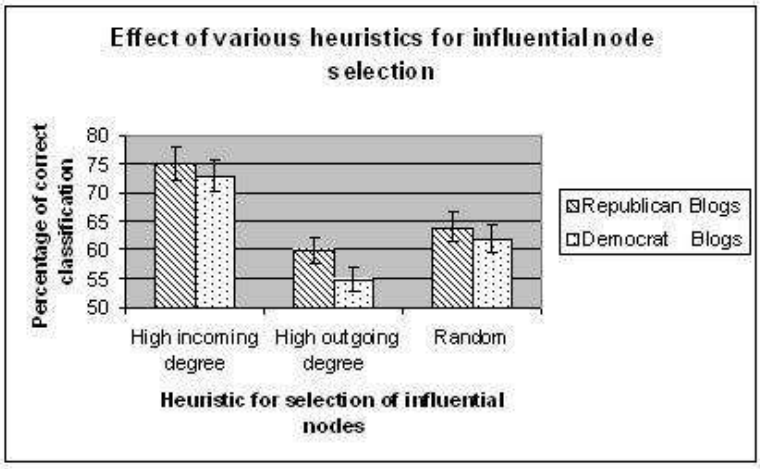

FIG. 4.6. High incoming degree as the heuristic for influential node selection gives best classification accuracy

incoming degree (or pagerank) is in fact the best heuristic. We do not have a very good insight on the plausible causes. The random selection demonstrates intuitive results. In order to ensure true randomness in the process of random selection, we selected 3 nodes at random, repeated this process 10 times and averaged the results. Also, we used "1111" as the bitset for atomic propagation.

4.5 Sample Polarity Computations

Table 4.1 depicts polarity values computed between some pairs of influential democrat and republican blogs. We present this data as a quick measure of demonstrating the potential of our work. We make the following interesting observations from 4.1.

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.

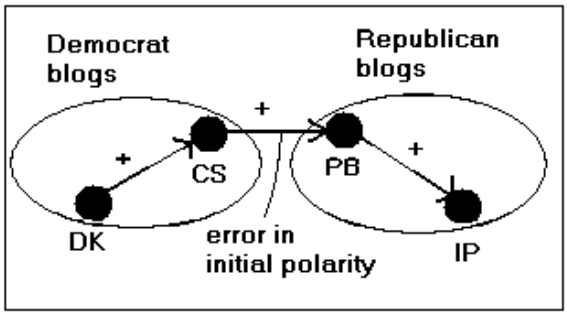

FIG. 4.7. Incorrect initial polarity computation on CS-PB link resulted in positive polar link between DK and IP. CS – http://crooksandliars.com, PB – http://powerlineblog.com

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. We analyzed the distribution of such values for some sample cases where a low polarity value resulted in generating incorrect polarity values for the blog nodes impacted by the incorrect polarity link. By *impacted nodes*, we mean direct neighbours of nodes having an incorrect polar link between them. Fig 4.7 presents a specific case of our analysis. Notice that we computed correct polarity for DK-CS and PB-IP links. However, the text surrounding the link where CS expressed an opinion about PB contained only one positive word “nice”. Since our lower threshold was set to *3*

Table 4.1. Polarity Values for Sample Influential Blogs

From-To	Number of links	Polarity before trust propagation	Polarity after trust propagation
MM–MM	0	N/A	3.53
MM–DK	0	N/A	-2.9
MM-IP	0	N/A	2.2
MM-AT	0	N/A	*1.09*
DK-MM	0	N/A	-2.9
DK-DK	0	N/A	2.02
DK-IP	0	N/A	*1.71*
DK-AT	20	0	8.51
IP-MM	8	1	2.2
IP-DK	6	0	*1.71*
IP-IP	0	N/A	1.06
IP-AT	0	N/A	-7.19
AT-MM	0	N/A	*1.09*
AT-DK	5	0.342	8.51
AT-IP	0	N/A	7.19
AT-AT	0	N/A	3.57

MM–http://michellemalkin.com, DK–http://dailykos.com
IP–http://instapundit.com, AT–http://atrios.blogspot.com

from our manual analysis, we assigned an incorrect value to this link. This incorrect value resulted in establishing a transitive connection between DK and IP after trust propagation leading to the wrong sign for polar link between them.

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.

4.6 Main Stream Media Classification

4.6.1 Motivation

Our test dataset from buzzmetrics [33] contained information about links from blog posts to main stream media sources like Washington Post[1], Boston Globe[2], MediaMatters[3] and such. As described in the previous sections, our experiments on determining the left or right inclination of blogs provide results with high acuracy. Hence, we decided to classify the main stream media sources using blog - media links. This serves as the evidence for the fact that our approach is not constrained to just the blogs - blog links but can be applied to other domains as well. Also, in the view of 2008 presidential elections, classification of main stream news sources has interesting business value.

4.6.2 Approach

Our approach for classification of main stream sources contains the same steps as described in chapter 3. Precisely, we compute the polarity for blog - media links and use the same trust propagation model to create a fully connected graph with polar links. The only difference in this experiment is that we consider a bi-partite graph of blogs and media sources. Thus, the graph does not contain any blog-blog links at bootstrap. One of the primary reasons of ignoring the blog-blog links is to avoid mis-classification due to instances of a “low-case” blog (blog having poor inlinks and outlinks) linking incorrectly to media sources and thus affecting the polarity from an influential blog to the media source under consideration. Since we do not have a labeled dataset of left and right leaning main stream sources, we do not validate our results formally. Instead, we used human subjects and resources from web to assess the quality of classification. This further required us to

[1] http://www.washingtonpost.com
[2] http://www.boston.com
[3] http://www.mediamatters.com

consider only the popular media sources for our experiments, so that our human experts could provide meaningful comments on the results. Thus, the size of graph (in terms of nodes as well as links) for this experiment is significantly lower than the precious experiments. For a graph of size comparable to the previous experiments, we believe that errors due to "low-case" blogs will be compensated in trust propagation and we need not make it bi-partite at bootstrap.

4.6.3 Results

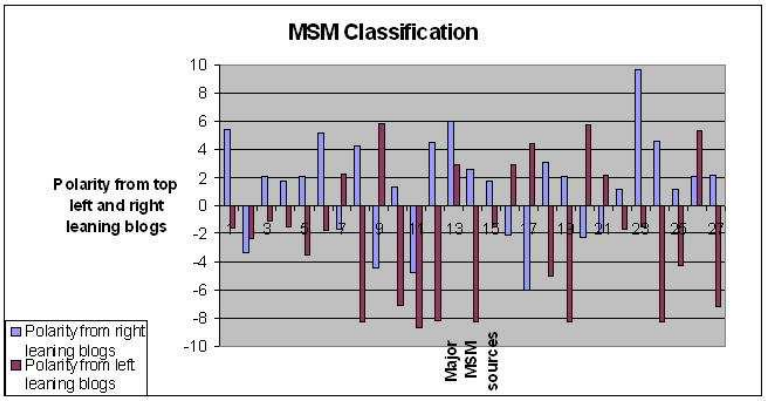

FIG. 4.8. Main Stream Media Sources can be classified correctly

Figure 4.8 represents the polarity values from influential republican and democrat blogs to media sources. The inclination of the media source can be interpreted from these results as follows:

If the polarity from republican blogs is positive and polarity from democrat blogs is negative, the media source is republican (right-leaning)

Else if the polarity from republican blogs is negative and polarity from democrat blogs is positive, the media source is democrat (left-leaning)

Else if the sign of polarity from both republican and democrat blogs is same, the media

1	http://www.washingtonpost.com	15	http://www.truthout.org
2	http://www.nytimes.com	16	http://today.reuters.com
3	http://news.yahoo.com	17	http://mediamatters.org
4	http://news.bbc.co.uk	18	http://www.townhall.com
5	http://www.msnbc.msn.com	19	http://www.timesonline.co.uk
6	http://www.cnn.com	20	http://www.guardian.co.uk
7	http://news.google.com	21	http://www.salon.com
8	http://www.usatoday.com	22	http://www.thenation.com
9	http://www.latimes.com	23	http://apnews.myway.com
10	http://www.boston.com	24	http://www.xaminr.com
11	http://www.abcnews.go.com	25	http://www.humaneventsonline.com
12	http://www.foxnews.com	26	http://www.dailybulltin.com
13	http://www.rawstory.com	27	http://www.spectator.org
14	http://www.cbsnews.com		

FIG. 4.9. Mapping from number to Main Stream Media Sources in 4.8

source is democrat or republican based on the respective magnitudes

We make the following interesting observations from the results presented in Figure 4.8

1. We can classify 24 out of 27 sources correctly.
2. Well-known left and right leaning sources like “guardian”, “foxnews”, “cnn”, “la-times”, “truthout” and “mediamatters” can be classified correctly.
3. The main outliers are “thenation” and “boston globe”.
4. “google news” is classified as left leaning, since the small snapshot in the dataset was indeed left leaning.
5. Both left and right leaning blogs talk negatively about “nytimes” and “abcnews” and positively about “rawstory” and “examiner”

Chapter 5

CONCLUSION AND FUTURE WORK

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 results are impressive. 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 good 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.

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 the topic of 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". We are also investigating new clustering techniques that incorporate polarity of links in the distance measure matrix and some of our preliminary results further confirm the effectiveness of link polarity. The idea of using link polarity suits well for all such domains where there exists a distinct set of different opinions (e.g. sports, windows vs. linux etc) and we believe that it has potential for deducing sub-communities from

communities as well.

While we are optimistic about our approach, we would like to note that the traditional clustering techniques [38] , [39] and [1] 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). The idea of link polarity can help in predicting the swings in such marginal nodes and the temporal analysis of such swings can be very beneficial for applications such as product advertising and viral marketing.

The main contribution of our work lies in applying trust propagation models over polar links. We believe that the idea of *polar links* very generic and can be applied to multiple domains. We demonstrated one such application in the domain of political blogosphere where we used natural language processing to deduce the link polarity. We would like to emphasize that the specific techniques to generate polar links is orthogonal to our main contribution. The idea of *Link Polarity* is very subjective to the domain under consideration. Hence in the discussion that follows, we give some insights into how our work can be extended to a very different domain of research conferences.

Co-authorship is an influential factor in the domain of research conferences. Suppose that the goal is to build a system for paper reviewers that assigns a *quality score* to the paper under review. Thus, the reviewer now has more metadata/feedback about the paper than just the contents of the paper. The system would be based on the data of papers, their authors and the affiliated universities from publicly available sources like DBLP[1] and citeseer[2]. The reviewer can assign trust/distrust (or *bias*) scores to the subset of researchers and universities that he is associated with. This score can serve as a measure of explicit user-driven “Link Polarity”. The system can use metadata such as how many times the author of the paper under review has published to a well-known conference, how respected

[1]http://www.informatik.uni-trier.de/ ley/db/

[2]http://citeseer.ist.psu.edu/

is the research community in the affiliated university and such, to generate more “polar” links. Using the trust propagation models discussed in our work, the system can then compute the trust/distrust score for the paper under consideration. This application can easily be extended to detect “conflict of interest” as well.

REFERENCES

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

[2] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 417–424, Association for Computational Linguistics, 2001.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, (Morristown, NJ, USA), pp. 79–86, Association for Computational Linguistics, 2002.

[4] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Comput. Linguist.*, vol. 16, no. 1, pp. 22–29, 1990.

[5] Das, , and C. M. Y., "Yahoo! for amazon: Sentiment extraction from small talk on the web," 2001.

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

[7] M. A. Hearst, "Direction-based text interpretation as an information access refinement," pp. 257–274, 1992.

[8] T. L., "Force dynamics in language and thought. in parasession on causatives and agentivity," 1985.

[9] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in *Proceedings of the ACL*, pp. 271–278, 2004.

[10] “Blogs as virtual communities: Identifying a sense of community in the julie/julia project,” *Into the Blogosphere: Rhetoric, Community and Culture of Weblogs*, 2004.

[11] L. Efimova and S. Hendrick, “In search for a virtual settlement: An exploration of weblog community boundaries.” 2005.

[12] S. C. Herring, I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu, “Conversations in the blogosphere: An analysis ”from the bottom up”,” in *HICSS ’05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS’05) - Track 4*, (Washington, DC, USA), p. 107.2, IEEE Computer Society, 2005.

[13] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, “Email as spectroscopy: automated discovery of community structure within organizations,” pp. 81–96, 2003.

[14] D. Fisher, “Using egocentric networks to understand communication,” *IEEE Internet Computing*, vol. 9, no. 5, pp. 20–28, 2005.

[15] J. Huang and M. S. Fox, “An ontology of trust: formal semantics and transitivity,” in *ICEC ’06: Proceedings of the 8th international conference on Electronic commerce*, (New York, NY, USA), pp. 259–270, ACM Press, 2006.

[16] Gans, M. Jarke, S. Kethers, and G. Lakemeyer, “Modeling the impact of trust and distrust in agent networks,” in *AOIS-01 at CAiSE-01*, 2001.

[17] T. Beth, M. Borcherding, and B. Klein, “Valuation of trust in open networks,” in *ESORICS ’94: Proceedings of the Third European Symposium on Research in Computer Security*, (London, UK), pp. 3–18, Springer-Verlag, 1994.

[18] M. Richardson, R. Agrawal, and P. Domingos, “Trust management for the semantic web.,” in *International Semantic Web Conference*, pp. 351–368, 2003.

[19] B. Yu and M. P. Singh, “Detecting deception in reputation management,” in *AAMAS ’03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, (New York, NY, USA), pp. 73–80, ACM Press, 2003.

[20] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, “The eigentrust algorithm for reputation management in p2p networks,” in *WWW ’03: Proceedings of the 12th international conference on World Wide Web*, (New York, NY, USA), pp. 640–651, ACM Press, 2003.

[21] 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*, (New York, NY, USA), pp. 403–412, ACM Press, 2004.

[22] D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins, “Information diffusion through blogspace,” *SIGKDD Explor. Newsl.*, vol. 6, no. 2, pp. 43–52, 2004.

[23] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose, “Implicit structure and the dynamics of blogspace,” May 2004.

[24] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo, “Deriving marketing intelligence from online discussion,” in *KDD ’05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, (New York, NY, USA), pp. 419–428, ACM Press, 2005.

[25] A. Qamra, B. Tseng, and E. Y. Chang, “Mining blog stories using community-based and temporal clustering,” in *CIKM ’06: Proceedings of the 15th ACM international conference on Information and knowledge management*, (New York, NY, USA), pp. 58–67, ACM Press, 2006.

[26] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, (New York, NY, USA), pp. 78–87, ACM Press, 2005.

[27] L. Lloyd, P. Kaulgud, and S. Skiena, "Newspapers vs. blogs: Who gets the scoop?," in *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[28] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "On the bursty evolution of blogspace," in *WWW '03: Proceedings of the 12th international conference on World Wide Web*, (New York, NY, USA), pp. 568–576, ACM Press, 2003.

[29] C. H. Brooks and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*, (New York, NY, USA), pp. 625–632, ACM Press, 2006.

[30] K. Shen and L. Wu, "Folksonomy as a complex network," 2005.

[31] C. Marlow, "Audience, structure and authority in the weblog community," May 2004.

[32] "Porter stemming algorithm." http://www.tartarus.org/ martin/PorterStemmer/.

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

[34] "Buzzmetrics dataset." http://www.icwsm.org/dataset.txt/.

[35] "International conference on weblogs and social media 2007." http://www.icwsm.org/.

[36] 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*, (New York, NY, USA), pp. 36–43, ACM Press, 2005.

[37] "Confusion matrix." http://en.wikipedia.org/wiki/Confusion_matrix.

[38] B. Tseng, J. Tatemura, and Y. Wu, "Tomographic clustering to visualize blog communities as mountain views," in *2rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th World Wide Web Conference*, 2005.

[39] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, p. 026113, 2004.

[40] J. Kleinberg, "Bursty and hierarchical structure in streams," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 91–101, ACM Press, 2002.

[41] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Trawling the web for emerging cyber-communities," in *WWW '99: Proceeding of the eighth international conference on World Wide Web*, (New York, NY, USA), pp. 1481–1493, Elsevier North-Holland, Inc., 1999.