Project Proposal
Regarding the Flow of Sentiment
Across Social Media Networks

Mahalia Miller
Conal Sathi
Daniel Wiesenthal

October 20, 2010

Abstract

This paper investigates the flow of various sentiment-oriented features through a hyperlinked network of blog posts or news media articles. In particular, this paper studies how the topology of the post cascade and the nature of the feature affect the flow behavior. We hope to determine certain lexical or topological features that have a high effect on whether and to what extent sentiment flows in a cascade.

Background

Blogs recently have become a popular medium for spreading news, thoughts, and commentary. With the rise of Web 2.0, it has become increasingly easier for someone to post his or her commentary on a recent issue. Blogs often cite and influence one another. It is our aim in this paper to examine the propagation of sentiment across a network of blogs.

Previous Work on Cascades

There has been previous work on analyzing linking patterns in large political blog graphs. Adamis and Glance (2005) examined different linking trends in Democratic and Republican blogs. They noted a divided blogosphere where liberal and conservative blogs tended to link to their own communities, rather than to each other. They also discovered the conservative blogosphere to be more densely connected and that the two communities tended to link to different news sources. There has also been work on analyzing patterns of cascading behavior in blog graphs. Lescovec et al (2007) discovered temporal patterns and topological patterns. They noted common cascade shapes, such as stars and chains, and cascade topological

properties, such as the size of cascades and their degree distribution. We will examine similar topological features when assessing how topology affects network flow of sentiment.

Previous Work on Sentiment Propagation

Though there is some literature on information diffusion in a network, tracking sentiment propagation in a network has not been explored in much depth. Nonetheless, Zafarani et al (2010) examined data from Live Journal, a social network where users maintain a blog. Users have the option of assigning a mood to each of their posts. The paper explores tracking the sentiment across the network by assigning each mood to a score and by defining mathematically if sentiment of one user has propagated to another user.

Zafarani et al will lay much groundwork in our project of tracking sentiment propagation across a network. Nonetheless, our paper departs from this paper in that we will use links between blog posts, rather than friendships on a social network to denote edges, and we will examine blogs from multiple sources, rather than just one source. Finally, we will use the content of the post to extract sentiment, rather than rely on the user to explicitly describe his sentiment.

Our Approach

This paper aims to understand the behavior of sentiment flow between blogs or news media articles. We will focus on cascades of blogs that are linked explictly by hyperlinks. The analysis will be divided into two categories: behavior of lexical features such as how happy a blog is and topological features, such as average degree of nodes in a cascade.

Data

This analysis will use the memetracker database provided by Prof. Leskovec. A difference over already-published results is that we will use new data that includes not only links, quotes, and data but also the full text of the post. We propose pruning the data in at least two ways. First, we will exclude Facebook and Twitter since we expect high amounts of noise in sentiment flow and very short texts. Second, we will look at non-trivial cascades by excluding singletons, which will greatly reduce the size of our data set. Leskovec et al (2007) suggest that over 98% of the posts are isolated. The details of the pruning methods will be described in the next section.

Methodology

Network Model

We will make our first pass over the Memetracker 2.0 data using the SNAP library. We chose this library because although the team is more familiar with NetworkX, we are worried that it would be prohibitively slow at importing so much data. We will import the data into a

graph as follows: each node will represent a blog post, and directed edges between nodes will represent hyperlinks. An edge will point from u to v where post u contains a hyperlink that cites post v. The weights associated with edges will be the time difference between the two posts. Additionally, each node will have as an attribute the full text of the post.

Data Refinement

After creating the graph, we will remove nodes from sources that we are not interested in– Facebook and Twitter, among (perhaps) others. We will also remove non-English language posts. After removing undesired nodes, we will extract cascade information in a similar fashion to Leskovec (2008) That is, we will find all cascade initiator nodes (nodes that have zero out-degree) and start following their in-links. Hopefully this will give us a directed acyclic graph with a single root node. We have not explored whether or not the data have updates to posts possibly linking other posts in the same cascade (which might create cycles), but if this issue arises we will develop a method to remove cycles. Again, as in Leskovec (2008), we will extract nodes that are part of multiple cascades multiple times so that each cascade can be examined in isolation. We will then prune our network of all singleton nodes, and, depending on the amount of remaining data, all cascades of maximum length less than some k.

Network Annotation

Once we have just the cascades we are interested in, we will attempt to freeze the network and move it from SNAP to NetworkX. (Again, this depends on an evaluation of the amount of data we have at this point in the process; if the dataset is still unwieldy, we may keep working with SNAP. Our preference, however, is to move to NetworkX since we would prefer to write our sentiment-extraction in Python.) Once the network is in NetworkX, we will annotate each node with values for a series of sentiment-oriented features. These features will include the overall ratio of positive to negative words as cross-referenced in the Harvard Inquirer dataset, the corresponding scores for positivity and negativity from the SentiWordNet dataset, and the average word objectivity score as calculated from the SentiWordNet dataset, among other things. Our goal is to create several sentiment features that can be extracted from just the full text of a single post. We will annotate each node in the network with the values its post attribute receives for these features. We will create features intended to capture sentiment splits along various axes, for example: positive/negative, objective/subjective, and supportive/unsupportive. We will also experiment with other features, such as sympathetic/unsympathetic, angry/joyous, optimistic/pessimistic, incitatory/calming, approach/withdrawal. We may also explore features focusing on the paradigm of plotting emotions on two axes of valence and arousal. Once this process has completed, we will annotate each node with baseline values for the features just mentioned. This will be done by looking at all nodes originating from the same blog source, and averaging the values for each feature over all posts of this blog at previous points in time.

Flow Analysis

With this fully annotated network, we will look at propagation of sentiment features by identifying the initial node in a cascade (not necessarily the root node) with feature(s) differing significantly from baseline and tracking the decay of that feature through the rest of the nodes in the cascade. An example output we might generate would be a graph plotting on the x-axis the progression through time (or number of nodes away from the root), and on the y-axis the level above baseline of a series of particular sentiment features (superimposed). We might hypothesize, for example, that objectivity would drop off steeply, as a sentiment feature that does not tend to flow well through networks, while negative emotion, on the other hand, might remain in the network through several nodes/timesteps.

Furthermore, we will examine network characteristics and analyze their effect on the flow of sentiment features. NetworkX has numerous built-in functions to aid the analysis. Specifically, we will graph the cascade length, average in-degrees, average out-degrees, average path length, and average degree, among other features, versus how well lexical features maintain their scores across a cascade. As in Leskovec (2008), we also will investigate cascade shapes, namely trees and stars, to catalog the types in our data set and see how the shapes might influence sentiment flow. The aim is to better understand the link between the topological features of the network and the propagation of sentiments.

Implications

An application of this understanding of flow is to create a classifier that can predict (above baseline accuracy) the sentiment information of an article. The features that we will put into this classifier will fall into two general categories: lexical, and topological. The lexical features will focus on extracting sentiment information from the textual content of a post. These lexical features will be used to describe articles. We will then use these lexical features to create higher-level topological features. Topological features will reflect the sentiment information present in the network structure in which the target node is integrated.

Possible challenges

A possible challenge will be if the noise in the data overpowers the signal and we are unable to see sentiment flow or deduce behavior. We hope to prevent this by pruning our data set to interesting cascades, as described above. We expect further NLP challenges such as ungrammatical posts or ones that confuse the classifiers with negation or sarcasm, but hope the data will nonetheless illustrate interesting flow behavior.

Conclusion

We will present results from our investigation in an academic paper of approximately 10

pages. The paper will address the behavior of sentiment flow in blog or news post cascades. Success will be defined by identifying distinct flow patterns for the different sentiment features and determining how topological features affect this flow. We hope that this work will provide the foundation for better understanding how a particular post will affect the sentiments of other posts that link to it.

References

[1] Adamic , L. and Glance, N. The political blogosphere and the 2004 U.S. election. Workshop on Link Discovery, 2005.

[2] Leskovec et al. Cascading behavior in large blog graphs: Patterns and a model. SIAM International Conference on Data Mining (SDM), 2007.

[3] Leskovec, J. Dynamics of large networks. PhD Dissertation, Machine Learning Department, Carnegie Mellon University, Technical report CMU-ML-08-111, 2008.

[4] Zafarani et al. Sentiment extraction in social networks: a case study in LiveJournal, In Advances in Social Computing , Vol. 6007, pp. 413-420, 2010.