Sentiment Propagation in Social Networks: A Case Study in LiveJournal

Reza Zafarani, William D. Cole, and Huan Liu

Computer Science and Engineering
Arizona State University
Tempe, AZ 85281-8809
{Reza,WCole,Huan.Liu}@asu.edu

Abstract. Social networking websites have facilitated a new style of communication through blogs, instant messaging, and various other techniques. Through collaboration, millions of users participate in millions of discussions every day. However, it is still difficult to determine the extent to which such discussions affect the emotions of the participants. We surmise that emotionally-oriented discussions may affect a given user's general emotional bent and be reflected in other discussions he or she may initiate or participate in. It is in this way that emotion (or sentiment) may propagate through a network. In this paper, we analyze sentiment propagation in social networks, review the importance and challenges of such a study, and provide methodologies for measuring this kind of propagation. A case study has been conducted on a large dataset gathered from the LiveJournal social network. Experimental results are promising in revealing some aspects of the sentiment propagation taking place in social networks.

1 Introduction

Social networks have become popular with the pervasive use of the World Wide Web. With the paradigm shift in the usage of the Web from information consumption to information production and sharing ("Web 2.0"), numerous social media services have emerged. Individuals use different social media services for various purposes and exhibit diverse behaviors. We use Flickr to share pictures with friends, Twitter to update our "status", MySpace to keep in touch with friends, and Blogs to express our interests, opinions, and thoughts. According to recent statistics[1], more than 10 billion photos exist on Facebook, 20 hours of video is uploaded on YouTube every minute, and around 38,400 photos are uploaded every hour on Flickr. With the massive amount of data published every day on these networks, we no longer have a shortage of experimental data; our challenge now is to make sense of the data. That said, the quantity of data gives us the opportunity to analyze the various behaviors of users in social networks and how they differ from their "real-world" social lives. Analogs of some

[1] http://www.labnol.org/internet/data-storage-for-user-generated-content/9656/

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 413–420, 2010.

real-world behaviors have been studied in the context of online social networks. For example, in [1], the authors introduce a technique to measure the degree of influence users in the Blogosphere have on other users in order to identify the most influential bloggers. In this study we focus on a different problem. Does the amount of [emotional] content users are exposed to on a daily basis in the online social world influence their emotions? And if it does, how can we observe this phenomenon? We aim to develop methodologies and find answers to these questions. In this paper, we present a case study with the following contributions:

- Formally define and study the propagation of sentiment in social networks,
- Quantify and predict the occurrence of a sentiment propagation, and
- Identify salient features that result in a sentiment propagation.

The rest of the paper is organized as follows: Section 2 describes the motivation behind this study. Section 3 presents the problem statement. Section 4 discusses a case study in LiveJournal, the approach used to analyze sentiment propagation in social networks, and experimental results. Section 5 summarizes the related research. Section 6 concludes with future work.

2 Motivations

The following five items describe the basic motivation behind our study.

- **How do individuals influence each other in social networks?** There is growing interest in the community to determine the extent to which participants can influence each other in terms of thoughts and behaviors via social networks.
- **Does sentiment propagate?** There has been extensive research on information diffusion in social networks. However, to the best of our knowledge this is the first study of its kind to consider sentiments as information entities to analyze the propagation thereof.
- **How does sentiment propagate?** Assuming there is a propagation effect, there are many questions that arise in this area. For example, how rapidly does sentiment propagate? What parameters influence the propagation rate? How do propagation speed variations correspond to real world events?
- **What different roles do individuals play in propagation?** It is important to analyze the actors involved in the propagation to understand these roles. For example, the users who initiate the propagation, those who relay the propagation (hubs), and those who block or enhance the propagation, are of interest in our a study.
- **How useful are sentiment analysis tools for sentiment propagation analysis?** It is interesting to find out how effective current sentiment analysis techniques are for analyzing sentiment propagation. For example, we used Normalized Google Distance (NGD) as measure of semantic distance in this study, and it is interesting to analyze how it affected our experimental results.

3 Problem Statement

Before we delve into the details, we will first formally define the problem of sentiment propagation in social networks. Let μ represent an active individual in cyberspace and s a single site. Without loss of generality, we restrict our study to a single website. We denote the set of all active users at site s as Λ_s. Let $m(\mu, t)$ denote the overall sentiment (mood) of user μ at time t. For a set of users $U \subset \Lambda_s$, we define the overall sentiment (mood) at time t as follows,

$$m(U,t) = \frac{\sum_{\mu \in U} m(\mu,t)}{|U|}. \tag{1}$$

Then, the sentiment propagation problem can be formally stated as follows:

Definition. ***Sentiment Propagation in a Social Network:*** *given a user μ, called the target user, a social network site s, and a subset of users $U \subset \Lambda_s$ that initiate the propagation at time t_i, a sentiment propagation has influenced μ at time t_j, iff.,*

$$|m(U,t_i) - m(\mu,t_j)| \leq |m(\Lambda_s,t_i) - m(\mu,t_j)| + b_1, \tag{2}$$

$$|m(U,t_i) - m(\mu,t_j)| \leq |m(U,t_i) - m(\mu,t_i)| + b_2, \tag{3}$$

where $t_i \leq t_j$ and $b_1, b_2 \geq 0$ are the intercepts.

Equation 2 denotes that if a propagation influences a user, then at time t_j, the overall mood of a user should be closer to the overall mood of the group that initiated the propagation than to that of the entire population of users. Equation 3, however, states that as time goes by (time t_j), the user should get even closer to the group that initiated the population in terms of mood. If both conditions are met, then we consider the user to be influenced by the propagation. Note that, b_1 and b_2 are parameters that can be learned or heuristically set.

4 A Case Study in LiveJournal

In order to analyze sentiment propagation, we performed a set of experiments on a dataset gathered in July 2009 from LiveJournal[2]. LiveJournal (LJ) is a web community where users can keep a blog or journal. Besides being quite popular, it includes features of a self-contained community and some social networking features similar to those of other social networking sites. An interesting feature of LiveJournal is that when users post on their blog they have the option of assigning a "mood" to their post. Users can select from a list of 130 moods or type freeform text to specify a mood not in the predefined list. Note that assigning moods to posts is optional and therefore not all posts have moods associated with them. We assume the mood assigned to a post is a sentiment that is generally oriented to the post content. This feature defines the data labels for our experiments.

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

Table 1. LiveJournal dataset statistics

Bloggers	Links	Link Density	Posts	Avg. Posts	Avg. Links	Net Diameter
16,444	131,846	9.8×10^{-4}	475,932	28.94	16	8

4.1 Data Collection

In order to gather a sufficiently large dataset for our experiments, we developed a breadth-first crawler that follows forward links (friendship relationship) on LiveJournal. We used forward links only since following them yields a large portion (or the "Weakly Connected Component") of the network [2]. The crawler starts from a well-known user found manually on LiveJournal, and reaches all users that are within four links of this user. We stored posts for every user found during the crawl. For each post, we stored the post date and the "mood" associated with it, if any. Overall, more than 1 GB of data was gathered. An overview of the dataset crawled is provided in Table 1.

4.2 Data Pre-processing

To ensure the data was sufficiently reliable for our experiments, various preprocessing steps were taken. First, we removed all non-English blog posts. We also removed all posts that did not have a mood associated with them (*non-moody* posts). We also filtered out posts with moods that were relatively infrequent in the corpus[3]. We then removed users who had less than five posts, had no *moody* posts (posts with moods), or had less than five friends. This was done to ensure we had a sample of user behavior sufficient enough to assess the sentimental dynamics which are the focus of this study. After these steps, the remaining posts in our dataset were associated with 285 distinct mood strings. We quantified the mood strings using Normalized Google Distance (NGD) [3]. Normalized Google Distance is a measure used for calculating the semantic distance between words from the hit counts returned by the Google search engine. Terms with similar meanings tend to be "close" in units of NGD, while dissimilar words tend to be farther apart. NGD between words x and y is defined as follows,

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}, \tag{4}$$

where M is the number pages on Google and $f(x)$, $f(y)$, and $f(x,y)$ are the number of hit counts for queries x, y, and $x + y$ on Google, respectively. In order to quantify moods, we first assumed that each mood can be represented using a pair $< positive, negative >$, where the first component is the positiveness and the second term is the negativeness. This approach has been extensively studied in opinion mining literature [4]. In order to obtain these values, a pair of *seed words* (poles) is required. Each seed word represents a pole (positive/negative). The NGD distance between these seed words and each mood

[3] In our experiment, the frequency threshold was empirically found and set to 10.

string provides us with the $< positive, negative >$ pair. Various seed words have been tested in the previous literature [5]. From these, we manually selected 2 pairs: $< excellent, poor >$ and $< happy, sad >$. This manual procedure involved looking at the NGD values produced for different moods using different seed words and ensuring that ostensibly negative (or positive) moods generated a greater negativeness (or positiveness) value. Then, all the mood strings were replaced with two pairs $< distance\ from\ happy, distance\ from\ sad >$, $< distance\ from\ excellent, distance\ from\ poor >$. After this step, the dataset was ready for experimentation.

4.3 Experiment Setup

Without loss of generality, we added a set of assumptions to the formalized version of the problem. First, we assumed that friends have greater influence over a user than non-friends. In other words, for a given user, the set of users that may initiate or transfer a sentiment propagation are his/her set of friends. Note that friends are directly connected to the user via forward links, i.e., are one hop away. The assumption simplifies the model but is still general enough to carry our experiments. Second, we assumed that the intercepts for our propagation constraints are zero, i.e., $b_1 = b_2 = 0$ (see Equations 2 and 3). Furthermore, we assumed that we have a propagation time window during which the sentiment propagation starts and ends. So, instead of arbitrary t_i's and t_j's (see Equations 2 and 3), we assumed that,

$$t_j - t_i \in \{1, 2, 3, 4, 5, 6\} \ \ months. \tag{5}$$

For each time period, we analyzed the dataset and checked if the conditions for a propagation are met for each user. We recorded the status of the propagation (i.e., a binary value with zero meaning the sentiment propagation did take place) in that period, along with additional data for each user: the number of posts made by the user in that period, the number of posts made by his/her friends in that period, the user's number of friends, the time window, and the total number of posts in that period. Again, this is a labeled dataset where *propagation status* is the class variable and all the other gathered attributes are its preliminary features that can be employed for classification. Next, we review our experimental results on this dataset, which revealed some aspects of the sentiment propagation taking place in social networks.

4.4 Evaluation Results

We first performed a propagation classification using the aforementioned dataset and 10-fold cross-validation to see how accurately we could predict sentiment propagation. Tables 2 and 3 show the classification results from C4.5, Sequential Minimal Optimization (SMO) [6], Naive Bayes (NB), Logistic Regression (LR), K^* (lazy learning), and Random Forest (RF), using a 10-fold cross validation and for both seed pairs used for the NGD metric. As shown in these tables, both decision tree algorithms (RF, C4.5) outperform other methods and

Table 2. Classification results for $< excellent, poor >$ seed words

	SMO	C4.5	NB	LR	RF	K*
Accuracy	62.65%	**71.10%**	61.22%	62.65%	**72.78%**	69.98%
Mean Absolute Error	0.37	**0.36**	0.45	0.46	**0.34**	0.38

Table 3. Classification results for $< happy, sad >$ seed words

	SMO	C4.5	NB	LR	RF	K*
Accuracy	69.51%	**76.16%**	67.95%	69.44%	**76.91%**	72.34%
Mean Absolute Error	0.30	**0.31**	0.40	0.42	**0.30**	0.35

result in reasonable accuracy in predicting the propagation based on available features. Moreover, it is evident that seed words play a role in the classification performance and should therefore be selected carefully.

Next, we analyzed the attributes of our dataset in order to find discriminatory ones. Table 4 shows the list of attributes and their average values for different class values, i.e., the binary variable that indicates whether or not a sentiment propagation exists. As shown in this table, users that exhibit sentiment propagation, on average, have more friends, make fewer posts, and have have less prolific friends than those of users who are not clearly subject to sentiment propagation. Moreover, propagation influences users in time periods shorter than 4 months. Note that the 4 month propagation window might be due to the nature of our dataset and it can be different in other datasets.

5 Related Work

Sentiment Analysis, also known as *Opinion Mining* or *Sentiment Extraction* is an emerging area of research as a subset of Text Mining. Sentiment Analysis is a newer research area at the crossroads of Text Mining and Computational Linguistics concerned not with the more commonly studied topical analysis of a document, but rather analyzing the overall polarity of opinions or sentiments expressed therein. It involves techniques to automatically analyze the sentiment, attitude, or opinion of textual documents on the World Wide Web, usually in terms of being positive, negative, or neutral. Generally speaking, it aims to determine the attitude of a speaker with respect to some topic. This attitude may indicate evaluational or affectational state (the emotional state of the author) or the intended emotional communication (the sentiment intended to be conveyed to the reader). Sentiment classification, as one of the active topics in this

Table 4. Average feature values for different classes

	# Friends	Time Window	Friend Posts	User Posts
Has Propagation	9.92	3.78	16.06	2.09
No Propagation	9.35	3.85	64.89	8.32

area, deals with categorizing sentiments. In most cases, if not all, the categories are limited to two (bipolar classification) [7]. These two categories represent the *positive* or *negative* sentiments. In other cases, these categories represent the objectivity/subjectivity of the textual excerpts [8,9]. A comprehensive list of computational measures regarding semantic relatedness for approximating the relative meaning of words/documents has been proposed in the literature. Some of these measures use lexical dictionaries such as WordNet, or SentiWordNet [10]. SentiWordNet provides 3 values for each sentiment (positive, negative, neutral). Other methods such as the Point-wise Mutual Information (PMI) [5] or the NGD [3] use search engines such as Google and Altavista to extract semantic similarity. Some early work in the study of information diffusion is [11], which introduced a model of collective behavior based on the concept of an aggregate threshold that must be overcome for individual behavior to spread to other actors. Another prominent model in this area, Independent Cascade Model (ICM), is alluded to in [12]. ICM models diffusion on a stochastic process whereby behavior spreads from one actor to another with a given probability. Some aspects of sentiment propagation has been analyzed before. For example, the intuition behind sentiment propagation in the 'real world' was validated in [13] through an analysis of responses to surveys given to a network of participants over a 20-year period. Moreover, Wu et al. devised a theory on the formation and spreading of opinions in a social network based explicitly on network structure, and make predictions about the pervasive influence of a minority of central actors [14]. Huberman et al. make the important observation that in computer social networks, significant influence occurs mostly between actors with a sufficiently close relationship [15]. This set of relationships is a subset of hard-linked relationships, i.e., 'friends', and is only identifiable by analysis of actual interactions between actors.

6 Conclusions and Future Work

In this paper, we studied the propagation of sentiments in social networks. We presented the challenges we encountered and developed approaches to tackle those challenges. We conducted carefully designed experiments on our dataset, gathered from LiveJournal, in order to analyze the sentiment propagation and to identify salient features that play a role in the propagation. Our preliminary experiments showed that users who have more friends, are less prolific, and who have friends who are less prolific exhibit greater sentiment propagation than users with dissimilar attributes. Also, propagation occurs within time periods of less than four months. In the future, we aim to expand our work by considering a user's overall sentiment as a time-dependent stochastic variable. This will address the shortcomings regarding the discretization of time in our current experiments. As mentioned earlier in our experiments, seed words play a role in the performance of detecting propagation. However, we did not attempt to find the optimum seed words in order to improve the accuracy of the classification methods. We leave this as another line of our future work, i.e., searching for an optimal seed set for improving accuracy.

Acknowledgements

This work is, in part, sponsored by AFOSR and ONR.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM 2008: Proceedings of the international conference on Web search and web data mining, pp. 207–218. ACM, New York (2008)
2. Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, p. 42. ACM, New York (2007)
3. Cilibrasi, R., Vitanyi, P., Cwi, A.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
4. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p. 354, Association for Computational Linguistics (2005)
5. Turney, P., et al.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
6. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. In: Advances in Kernel Methods-Support Vector Learning, vol. 208 (1999)
7. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86. Association for Computational Linguistics, Morristown (2002)
8. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 486–497. Springer, Heidelberg (2005)
9. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), pp. 105–112 (2003)
10. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC, Citeseer, vol. 6 (2006)
11. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology 83(6), 1420–1443 (1978)
12. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137–146. ACM, New York (2003)
13. Fowler, J., Christakis, N.: Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. British Medical Journal 337(dec04 2), a2338 (2008)
14. Wu, F., Huberman, B., Adamic, L., Tyler, J.: Information flow in social groups. Physica A: Statistical Mechanics and its Applications 337(1-2), 327–335 (2004)
15. Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (2008)