Understanding Retweeting Behaviors in Social Networks

Zi Yang[⋆†], Jingyi Guo[⋆], Keke Cai[♮], Jie Tang[⋆†], Juanzi Li[⋆†], Li Zhang[♮] and Zhong Su[♮]
[⋆] Tsinghua National Laboratory for Information Science and Technology
[†]Department of Computer Science and Technology, Tsinghua University
[♮] IBM, China Research Lab
{yangzi, tangjie, ljz}@keg.cs.tsinghua.edu.cn, {caikeke, lizhang, suzhong}@cn.ibm.com

ABSTRACT

Retweeting is an important action (behavior) on Twitter, indicating the behavior that users re-post microblogs of their friends. While much work has been conducted for mining textual content that users generate or analyzing the social network structure, few publications systematically study the underlying mechanism of the retweeting behaviors. In this paper, we perform an interesting analysis for the problem on Twitter. We have found that almost 25.5% of the tweets posted by users are actually retweeted from friends' blog spaces. Our investigation unveils that for the retweet behaviors, some statistics still follows the power law distribution, while some others violate the state-of-the-art distribution for Web. Based on these important observations, we propose a factor graph model to predict users' retweeting behaviors. Experimental results on the Twitter data set show that our method can achieve a precision of 28.81% and recall of 37.33% for prediction of the retweet behaviors.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data Mining; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms

Algorithms, Experimentation

Keywords

Twitter, Retweet behavior, Social influence, Factor graph

1. INTRODUCTION

The rapidly developing Web-based social applications and media, such as Facebook, Twitter, and Flickr, have attracted much attention. Message forwarding (e.g., retweeting on Twitter.com) is one of the most popular functions in many existing social networks. In twitter, people can choose to retweet messages on their blog space. In this way, the information carried by the message can be quickly spread in the social network. A simple example is shown below, when a user is viewing the message m:

$$\text{RT @}\underbrace{\text{dahara}}_{t_{m,2}}\text{RT @}\underbrace{\text{Three\_Ten}}_{t_{m,1}}\underbrace{\text{Stunning Pictures of Strange}\ldots}_{c_m}$$

which indicates that a message was originally posted by "Three_Ten" and was retweeted by "dahara", and now is retweeted by the current user again. In general, a retweeted message consists of *trace* and *content*. Contents that we refer to can be extended to any forms as long as it can be forwarded from users to others or shared by users with all their friends, such as real blogs, photos and external links.

While much work has been conducted for mining textual content that users generate or analyzing the social network structure [4, 2, 5, 7], few publications systematically study the underlying mechanism of the retweeting behaviors [1, 8]. Some specific messages posted by particular users are more likely spread widely while others attract attention from few users. It should be said that various reasons make the message to be propagated, and it is interesting to investigate why some messages can be spread widely in the social network. Specifically, we analyze the retweeting behavior for each individual user and message, and aim to understand why some users tend to retweet messages, while others not, and what factors are responsible for acts of retweeting of messages. Accordingly, an interesting and fundamental question is: can we predict the retweeting behaviors based on users' history behaviors and the (global or local) trend on the Web?

2. PROBLEM DEFINITION

Based on the analysis of the factors that are attributed to users' decisions, messages, users, and the relations between them will be formally defined in this section, and then we formalize this problem.

To model the retweeting behavior, we collect a set of instances that describe the scenario when u_i receives a *message* $m_{ij} = \{c_j, \mathbf{p}_{ij}\}$ at time t_{ij}. Variable c_j denotes the content of the original message m_{ij}. Vector $\mathbf{p}_{ij} = \{p_{ij1}, p_{ij2}, \ldots, p_{ijl_{ij}}\}$ denotes the trace of the message seen by the user. Although the contents seen by different users may be the same, the traces could be different among users. $p_{ij1} \in U$ is the initial poster of the message, $p_{ijk} \in U$ is the k-th bearer of the message m_{ij} on the trace, and l_{ij} is the length of the trace. The follower-followee relationships are required to be satisfied between each consecutive user pair on the trace.

CIKM'10, October 25–30, 2010, Toronto, Ontario, Canada.

The spreads of messages proceed in a cascading fashion successively from the user who posts the message to some of his/her followers, and subsequently to some followers' followers. Thus, each message in the input set of instances has an ancestor message, by linking all the retweets with their ancestors, we could obtain a set of retweet threads. in other words, retweeting instances constitute a set of directed *retweeting trees* embedded in the friendship network.

Definition 1. RETWEETING PREDICTION PROBLEM: Given a social network G and a set M of tweets and retweet behaviors in history, we aim to predict (1) if users will retweet the tweet m to their friends after viewing it, (2) the range of spread for a new tweet m written by user u.

The first subproblem (referred to as *local prediction*) aims at predicting the behavior of users when a message posted by one of their friends has already appeared on their timeline, in other words, we assume that the user has a probability (> 0) in reading and retweeting the message. However in a more flexible conditions instead of being given a complete scenario, more factors are responsible for their retweeting/ignoring behaviors, especially the behaviors of their friends are highly related with their behaviors, which is depicted in the second problem (referred to as *global prediction*).

3. MODELING RETWEET BEHAVIOR

In this section, we formalize our problem in a semi-supervised learning framework. It is then tackled by a factor graph model [3, 6, 9] that incorporates the local prediction problem with the global constraint. An approach based on max-sum algorithm is applied to train the probabilistic model.

3.1 Basic Idea

As we describe in Section 2, the positive instances (labeled with $y_{ij} = 1$) constitute a set of retweeting trees, and by integrating the negative instances (labeled with $y_{ij} = 0$), we augment the retweeting trees into a set of *augmented retweeting trees* $\{A_d\}_d$ (ARTs). The root of each ART represents the initial poster of the message, the interior nodes represent the positive instances, corresponding to the retweeting behaviors, whereas the leaves represent the negative instances, corresponding to the ignoring behaviors.

The problem can be formalized as a semi-supervised learning problem. We collect a set of labeled instances T consisting of non-root nodes $\{\langle (u_i, c_j, \mathbf{p}_{ij}, t_{ij}), y_{ij} \rangle\}_{i,j}$ of complete ARTs, and a set of unlabeled instances S consisting of instances from a series of incomplete ARTs. When we try to estimate the propagation of a certain newly created message, the content c_j and the initial poster p_{ij1} are known for each user $u_i \in U$, and we may assume the content does not change during the propagation, however, the length of the trace l_{ij} and all the other trace users $p_{ij1}, \dots, p_{ijl_j}$ are unknown. Therefore, the content c_j, the initial poster p_{ij1} and the elapsed time t of the instances in T are fixed, whereas we could enumerate all the possible traces $\mathbf{p}_{ij}$ starting from p_{ij1} and ending at some users $\{u_i\}_i$. Such a series of ARTs constitute the solution space, and finally, the unknown labels y_{ij} is determined.

3.2 Definition of Factors

We introduce a factor f_{ij} that ensures the extracted features $\mathbf{x}_{ij} = (x_{ij}^1, \dots, x_{ij}^{22})^\top$ could represent the decision y_{ij}. We also introduce a set of weight variables, denoted by $\Lambda = (\lambda^k)_k^\top$ for each extracted feature. Each feature factor f_{ij}^k is defined as a regularized sum-of-square error function as follows:

$$f_{ij}(\Lambda) = \exp\left[-\left(y_{ij} - \Lambda^\top \mathbf{x}_{ij}\right)^2 - \beta \Lambda^\top \Lambda\right] \tag{1}$$

where β is the regularization coefficient that controls the relative importance of the data-dependent error and the regularization term $\Lambda^\top \Lambda$.

For unlabeled instances, we could define a feature factor $g_{ij}(y_{ij}, \Lambda)$ similarly as for the labeled instances. Since the calculation for the features depends on the choice of trace $\mathbf{p}_{ij}$, and subsequently the factor should incorporate the unknown variable, i.e.,

$$g_{ij}(y_{ij}, \Lambda, \mathbf{p}_{ij}) = \exp\left[-\left(y_{ij} - \Lambda^\top \mathbf{x}_{ij}(\mathbf{p}_{ij})\right)^2 - \beta \Lambda^\top \Lambda\right] \tag{2}$$

From the perspective of validity of retweeting trees, we introduce a constraint factor for each unlabeled instance, and it is also defined in terms of y_{ij}, which verifies the validity of each propagation path from the initial poster p_{ij1} to the current user u_i. Equivalently, the constraint can be defined for each consecutive bearers, e.g., u_i and the preceding user p_{ijl_j} (denoted as u_{i_I}) for the message m_{ij} instead of the current user u_i and the whole trace $\mathbf{p}_{ij}$, and formally,

$$h_{ij}(y_{ij}, y_{i_I j}) = \begin{cases} 0 & \text{if } y_{ij} \wedge \neg y_{i_I j} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

To make the constraint satisfied, we could alter the values of the decision variables y_{ij} in two different ways: by setting $y_{ij} = 0$ or setting $y_{i_I j} = 1$. Both approaches bring effects differently on other factors, the gains and losses of both approaches are evaluated in the optimization of the global objective by our algorithm.

In sum, the solution for problem that defined in Section 2 is tackled by solving the optimization objective, i.e., the normalized product of Eq. 1, 2, 3.

$$\max_{\Theta} \frac{1}{Z} f_{ij}(\Lambda) \prod_{(i,j)\in T} \left[g_{ij}(y_{ij}, \Lambda, \mathbf{p}_{ij}) \cdot h_{ij}(y_{ij}, y_{i_I j})\right] \tag{4}$$

where Z is a normalization factor, Θ is the set of variables, which includes $\{y_{ij}\}_{(i,j)\in S}$, Λ, $\{u_{i_I}\}_{(i,j)\in S}$. $\{\mathbf{p}_{ij}\}_{(i,j)\in S}$ could be determined by $\{u_{i_I}\}_{(i,j)\in S}$.

We derive an iterative algorithm for the above objective function based on the loopy max-sum algorithm. Details are omitted due to space limitation.

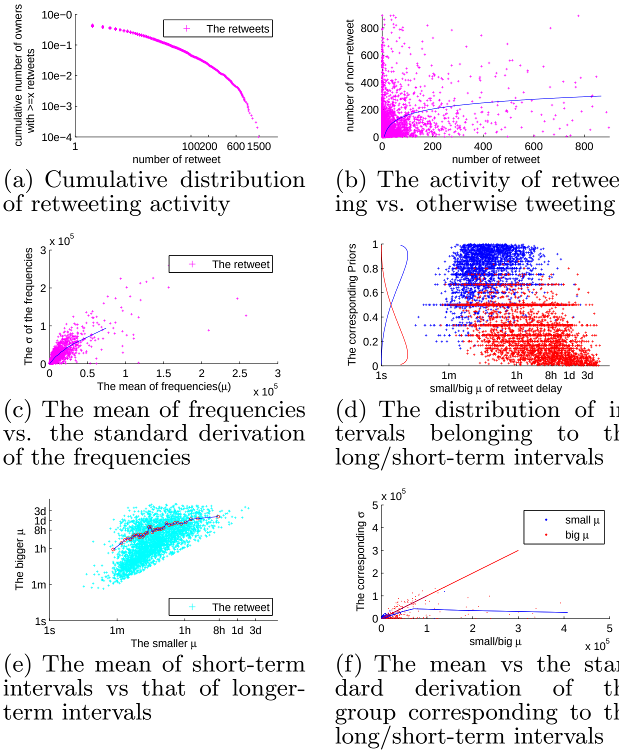
4. EXPERIMENT

In this section, we first statistically study the retweeting data, and then define features for modeling retweet behaviors. We conduct experiments on the Twitter dataset for retweet prediction and spread prediction.

4.1 Observation

In this section, we demonstrate a statistical analysis on some specific features that motivate users to retweet instead of acts of ignoring messages, and conclude notable results.

Retweeting activity of users For each retweet, we count the total number of retweets from each retweeter within the period of time and sort them in descending order with fixed



(a) Cumulative distribution of retweeting activity

(b) The activity of retweeting vs. otherwise tweeting

(c) The mean of frequencies vs. the standard derivation of the frequencies

(d) The distribution of intervals belonging to the long/short-term intervals

(e) The mean of short-term intervals vs that of longer-term intervals

(f) The mean vs the standard derivation of the group corresponding to the long/short-term intervals

Figure 1: Retweeting activity of user

unit interval. From Figure 1(a), we can see that most of the users retweet at a low frequency (and the average number of retweets of a user within 7 days is 197) and only a few users are retweet-aholic. A portion of 3.13% of the retweets are posted by users who retweet more than 1,000 times. It satisfies the long tail theory, i.e., the activity behavior of each individual user cannot be inferred from the global retweeting behavior, which motivates us to further analyze what factors are actually attributed to the diversity of activity in retweeting.

The activity of retweeting vs. tweeting. To conclude the users' activities of retweets from their local behaviors, e.g., the tweeting behavior, we may first ask whether it is highly related with their behaviors of writing tweets. Thus in Figure 1(b), we plot the number of the retweets against the number of the non-retweets (including normal tweets and replies) for each user in pink, with a 50-point moving average indicated in blue. We see that many users rarely retweet messages but post many other kinds of tweets, corresponding to the segment on the curve above the diagonal line when the number of retweets < 200, and many other retweet-aholics post much fewer tweets than retweets.

Analyze users' regularity of retweeting. For each user who retweets more than twice in the period, we make statistics of the distribution of time periods between two consecutive tweets, and the result is shown in Figure 1(c). The x-coordinate of a point describes the mean of the user's tweeting intervals, and the y-coordinate describes the standard deviation. The high overlap of the fitted line with the diagonal line indicates that the regularity is scale-invariant, i.e., when users become more tweet-aholic, the standard derivations of the intervals increase in the same proportion.

In fact, there may be two types of intervals that contribute to the deviation: some are the actual tweeting intervals between two tweeting behaviors, and the others are long-term pauses between periods of intensive tweeting actions, (e.g., due to daily sleeping routines). These pauses take up a small portion of the intervals, but contribute significantly to the calculation of standard deviation, and further distort the accuracy of discovered patterns of regularity. We simply assume this phenomenon is generated from a mixture of two Gaussian distributions. For each user, we plot the number of users' tweeting intervals that belong to the short-term intervals/longer-term intervals in Figure 1(d) in x-logarithmic scale with fitted Beta distributions of the marginal probabilities over them in blue, two means in Figure 1(e). We can see that the average curve consists of two segments. The first segment ends when the short-term interval reaches half an hour (corresponding to tweet-aholics), which is almost the duration of off-Internet for normal users, whereas the long-term intervals are prolonged with the tweeting frequency decreasing. We also plot the mean against the SD of both kinds in Figure 1(f). We can see that the mean regularities for both short-term and long-term intervals for users to tweet could be bounded.

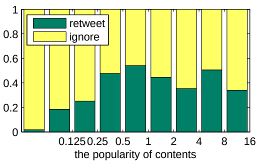


Figure 2: Importance of content

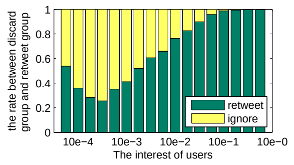


Figure 3: Interest of user

Importance of content. The importance of content can be estimated from the frequencies of the terms used by all the messages of all the users within this period of time. It is motivated by the fact that a hot topic is discussed by a relative majority of the community and the topics of special interests are known by few users. Hence, to describe the importance of content, we calculate the sum of tf-idf values of keywords in the content of the original message, i.e.,

$$\text{tf-idf}(m) = \sum_t n(t, m) \cdot \log \frac{|M|}{n(t, M)} \tag{5}$$

where $n(t, m)$ is the number of occurrences of the term t in message m, $n(t, M)$ is the number of occurrences of the term t in the whole message collection M, and $|M|$ is the total number of messages. It is compared with the importances (tf-idf values) of the ignored tweets. The result is shown in Figure 2 in x-logarithmic scale.

Note that the same original message may be retweeted by some of their followers, and ignored by some others. The content of the message is taken into account in both situations (retweeting and ignoring). From Figure 2, we can see that if tf-idf value is greater than 0.25, the messages will more likely be retweeted than those with importance scores less than 0.25. We can also see that in all the cases the length of trace is shorter than six, the importance of content exhibits almost the same. Since the total number of messages with trace length greater than six is limited, there is little statistical significance.

Interest of user. To jointly consider the mutual correlation between the user and the content of the message, we estimate how much the user is interested in the content of the message. The Jaccard distance is applied to calculate the similarity between the user u (represented by the bag of terms used in all his tweets) and the original message m as follows: $\frac{|u \cap m|}{|u \cup m|}$. A higher overlap of the two sets of terms implies that a higher probability that the user is interested in the content of message. The distribution of all the similarity scores is shown in Figure 3 in x-logarithmic scale.

We can see from the figure that if the users are interested in the message (the Jaccard distance is greater than 10^{-2}), there is a higher probability for them to retweet it. If there are rarely common terms, then there is still a relatively high probability to retweet. Comparatively, if Jaccard distance is among 10^{-4} and 10^{-3}, users are more likely to ignore them. In fact, the average similarity of an arbitrary tweet and an arbitrary user lies in the same interval.

4.2 Settings

Feature Description. As we have shown in Section 4.1, the delay periods are located around two different means (corresponding to short-term period t_i^{short} and long-term t_i^{long} period). To predict the willingness of retweeting at time t_{ij}, $\text{delay}(u_i, t_{ij})$ is defined as

$$\begin{aligned}&\text{delay}(u_i, t_{ij})\\&= \exp\left(-n_i^{\text{short}}(t_{ij} - t_i^{\text{short}})^2 - n_i^{\text{long}}(t_{ij} - t_i^{\text{long}})^2\right)\end{aligned} \tag{6}$$

where n^{short} and n^{long} are the prior probabilities calculated with EM algorithm of Gaussian mixture model. Intuitively, we could expect that the user u_i with higher probability responses to retweet at two different latencies corresponding to t_i^{short} and t_i^{long}. We define 22 features in the training process extracted from the users' history preferences, messages' contents, information of the trace, and the time delay and as the basis for the further prediction. Details are omitted due to space limitation.

Dataset. In the task of predicting the pair-wise retweeting behavior, we split the datasets on the level of ARTs, instead of instances, into training set and testing set. To evaluate the propagation of messages, we choose a sub dataset from that used in predicting the pair-wise retweeting behavior.

Measures and Baselines. For evaluating the prediction of pair-wise retweeting behavior, *precision* and *recall* are used to evaluate the performance. For evaluating the prediction of propagation spread of messages, it can be considered as a "retweeter retrieval" task, and both concepts in IR scenarios are applied. We apply two state-of-the-art methods in classification tasks, linear SVM and L1-regularized logistic regression (LogReg) for labeling the instances.

4.3 Results of Prediction

By employing the method proposed in Section 3 and the results are shown in Table 4.3. We can see that the proposed method does not outperform the baseline methods for the pair-wise classification task. In fact, as a task that only cares the isolated features of instances, it has been demonstrated that SVM and LogReg are more capable of representing the characteristics of the distribution of samples in feature space, and tend to be more discriminable for unseen data. The discriminability of our approach is only based on the factor f, which is the least sum-of-square errors.

For the task of predicting the spread of messages, we iteratively run the SVM classifier and the LogReg classifier for determining the behavior of a user that is consecutive to the initial poster or a "retweeter" that has been predicted to retweet in a cascade fashion. Different from the baseline methods, our method considers the entire graph as a whole, and the behaviors of users will be automatically judged by not only their followees, but also followers, who still provide valuable information for the consistency of the complete propagation. Our method can achieve a performance of 0.2881 in terms of precision on average, whereas both SVM and LogReg fail to predict the range of propagation of messages, with precisions of 0.0144 and 0.0052 respectively. Our method outperforms LogReg but fails for SVM in terms of recall, which is because the recall is the arithmetic average for all the instances, and our method predicts that the message of some "short" ARTs will not be propagated.

Table 1: Results of prediction

	Method	Precision	Recall	F-measure
Pair-wise behavior	Our approach	0.8003	0.6242	0.7014
	SVM	**0.8437**	**0.7358**	**0.8141**
	LogReg	0.8115	0.7131	0.7899
Message spread	Our approach	**0.2881**	0.3733	**0.3252**
	SVM	0.0144	**0.6084**	0.0281
	LogReg	0.0052	0.3047	0.0102

5. CONCLUSION

In this paper, we perform an interesting analysis for the retweet problem on Twitter, and discover some interesting phenomena. Specifically, we analyze how the retweet behavior is influenced by factors: user, message, time, etc. Based on these important observations, we propose a semi-supervised framework on a factor graph model to predict users' retweet behaviors. Experimental results on a data set show that our method can achieve an precision of 28.81% and recall of 37.33% for prediction of the retweet behaviors in completely actual scenarios.

6. *ACKNOWLEDGMENTS

The work is supported by the Natural Science Foundation of China (No. 60703059, No. 60973102), Chinese National Key Foundation Research (No. 60933013), National High-tech R&D Program (No. 2009AA01Z138).

7. REFERENCES

[1] d. boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS '10*, 2010.

[2] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *KDD '08*, 2008.

[3] F. Kschischang, S. Member, B. J. Frey, and H. andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 2001.

[4] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM '07*, 2007.

[5] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *PNAS*, 105(12):4633–4638, 25 March 2008.

[6] H. Loeliger. An introduction to factor graphs. *IEEE Signal Processing Magazine*, 21(1):28–41, January 2004.

[7] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD '09*, 2009.

[8] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. *ICWSM '10*, 2010.

[9] Z. Yang, J. Tang, and J. Li. Social community analysis via factor graph model. *IEEE Intelligent Systems*, 2010. (To appear)