Predicting Popular Messages in Twitter

Liangjie Hong Ovidiu Dan Brian D. Davison
Dept. of Computer Science and Engineering, Lehigh University
Bethlehem, PA 18015 USA
{lih307, ovd209, davison}@cse.lehigh.edu

ABSTRACT

Social network services have become a viable source of information for users. In Twitter, information deemed important by the community propagates through retweets. Studying the characteristics of such popular messages is important for a number of tasks, such as breaking news detection, personalized message recommendation, viral marketing and others. This paper investigates the problem of predicting the popularity of messages as measured by the number of future retweets and sheds some light on what kinds of factors influence information propagation in Twitter. We formulate the task into a classification problem and study two of its variants by investigating a wide spectrum of features based on the content of the messages, temporal information, metadata of messages and users, as well as structural properties of the users' social graph on a large scale dataset. We show that our method can successfully predict messages which will attract thousands of retweets with good performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design, Experimentation

Keywords

Information Diffusion, Social Media, Classification, Microblogs

1. INTRODUCTION

Social network services such as Facebook, Myspace and Twitter have become important communication tools for many online users. Such websites are increasingly used for communicating breaking news, eyewitness accounts and organizing groups of people. Users of these services have become accustomed to receiving timely updates on important events, both of personal and global importance. However, the flow of information within the social graph can lead to two problems. First, even users with few friends can experience information overload due to the high volume of messages. Second, users with a small number of social connections might miss important messages which do not reach them. In this second situation it would be useful to recommend interesting messages to such users, which might lead them to follow new users and augment their social network. Both problems require us to determine the importance of messages. In this work we use "retweets" as a measure of popularity and address the problem by utilizing machine learning techniques to predict whether and how often new messages will be retweeted in the future. We treat the problem as a classification task. First, we train a binary classifier with positive and negative examples of messages which will be retweeted in the future. Second, we train a multi-class classifier which predicts the volume range of future retweets for a new message. To build the classifiers, we investigate a wide spectrum of features to determine which ones can be successfully used as predictors of popularity, including the content and topical information of messages, graph structural properties of users, temporal dynamics of retweet chains and meta-information of users. We conduct our experiments on a large scale dataset. The results suggest that we can successfully predict whether a message will be retweeted or not. Furthermore, we can also predict the volume of retweets with good predictive performance. Similar work by Suh et al. [2] studied a variety of factors that might influence retweets without explicitly showing their effectiveness in a classification framework.

Table 1: Sample tweets with high retweets

RT @paramore Watch the World Premiere of Paramore's new video for 'Brick By Boring Brick' #paramore
RT @CamaroWRX: http://bit.ly/794Edz because everyone #needsmorebradley
RT @narendra: Please RT. Some recent thoughts on the empathic web. that made the Huffington Post - http://bit.ly/9WyxnT

2. PROBLEMS & FEATURES

We cast the problem of predicting the popularity of messages into two classification problems: 1) a binary classification problem that predicts whether or not a message will be retweeted, and, 2) a multi-class classification problem that predicts the volume of retweets a particular message will receive in the near future. Two messages are considered identical if they share the same MD5 value. We sort all such messages by ascending time order, forming a chain of messages. The first message in the chain is the earliest version of the message in our dataset and all later messages in the same chain should contain at least one "RT" term. For the binary classification problem, for n messages in a chain, the first $n-1$ messages are considered as "positive instances" and the last one as a "negative instance". In addition, all other messages which are not in any chains are also considered as "negative instances". For the multi-class classification problem, the number of messages following a specific message is treated as the number of retweets this message will receive. Therefore, for all messages, we assign a non-negative integer value to represent the number of subsequent retweets. It is difficult to predict the exact number of retweets a

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

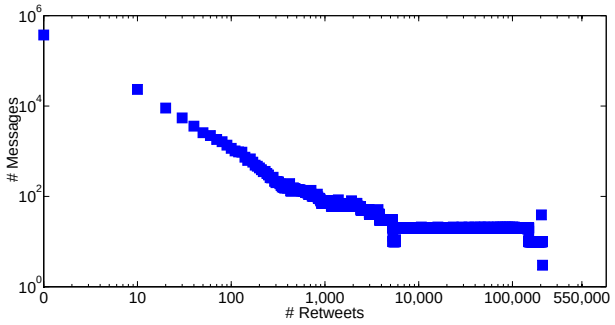

Figure 1: Retweet distribution

Table 2: The performance of two classification tasks

Methods	Precision	Recall	F_1
Retweet Before	0.685	0.166	0.263
TF-IDF	0.310	0.249	0.276
Our Method	0.993	0.435	0.603
Without *User Retweet*	0.993	0.399	0.569
Without *Degree Distr.*	0.473	0.470	0.471
Without *Retweet Before*	0.678	0.250	0.364
	Previous Method	**Temporal**	**User Activities**
Class	**Accu.**	**Accu.**	**Accu.**
0	0.9999	0.9999	0.9999
1	0.0338	0.1139	0.1490
2	0.0522	0.4309	0.4346
3	0.9896	0.9004	0.9209

particular message will receive due to the fact that the maximum volume of retweets in the test set may be much higher than in its training set. Therefore, rather than directly predicting these integer values, we relax the problem by defining several categories to represent the volume of retweets and predict these categories instead. We investigate a wide range of possible features including content features, graph topological features, temporal features and metadata features, all of which are applicable to the two classification tasks. For content features, we use TF-IDF scores as a baseline and utilize Latent Dirichlet Allocation (LDA) to obtain the topic distributions for each message. For topological features, multiple popular features are considered, including global and local structures, such as PageRank, degree distribution, local clustering coefficient and reciprocal links. We also assume that users may track hot topics and may quickly switch topics over time. Therefore, for temporal features, we measure the time difference between the current message and the origin within the chain, the time difference between the current and the previous tweet, the average time difference of consecutive messages in the same chain and the average time a user's messages get retweeted. For meta information features, we are mainly interested in whether a message has been retweeted before (Retweet Before) and how many times in the past the messages generated by a particular user have been retweeted (User Retweet). Other user activities such as the total number of messages a user produced are also considered.

3. EXPERIMENTS

We run our experiments with messages collected in November and December 2009, as well as the immediate social graph of the users which are active in this time period. The dataset contains 10,612,601 messages and 2,541,178 users. For the binary classification problem, we report Precision, Recall and F_1 score. For the multi-class classification problem, we report accuracy for each class. Instead of directly predicting the exact number of retweets, we divide the messages into different retweet volume "classes": 0: Not retweeted, 1: Retweeted less than 100, 2: Retweeted less than 10000, 3: Retweeted more than 10000, similarly to [1]. We binned messages with retweets (size equals 10) and plot the distribution in Figure 1, showing that a large number of messages do not receive any significant retweets (less than 10) and a significant amount of messages also have more than 10000 retweets. Three messages with most retweets in Class 3 are shown in Table 1 as examples where the first one is about entertainment, the second is related to hockey and the third one is related to IT. We use Logistic Regression as the classifier, as we found it to be the most stable in preliminary experiments. We train the classifier on one week and test it on the next week and the results are the averages over four such experiments.

We compare the results to two obvious baselines. One is based on whether a messages has been retweeted or not in the past, and the other uses the TF-IDF scores for terms in each message. The results are shown in the upper part of Table 2. For both baselines, the F_1 score is roughly around 0.27, with relatively low Precision and Recall scores. For our method, the best performance is achieved by using "TF-IDF", "LDA", "degree distribution", "Retweet Before" and "User Retweet" features while all measures are much better than both baselines. In order to investigate the effectiveness of each feature, we remove individual features and see how the performance is affected. This is shown in the middle of Table 2, where the first column shows which feature have been taken away. We observe that the performance drops significantly when "degree distribution" and "Retweet Before" are removed, implying that they contribute greatly to the overall performance. We conjecture that in Twitter, users mainly see messages from their first level friends whom they are directly following. Therefore, the likelihood that some of a user's followers will retweet a new message depends on how many followers the user has. Thus, the effect of in-degree is obvious to some extent. For out-degree, since many messages which will be retweeted in the future have already been retweeted, a user with limited friends may face difficulties in seeing such messages. For multi-class classification, the lower part of Table 2 shows the performance achieved by using the best features in the binary classification. The results are extremely good in Class 0 and Class 3 but poor in Class 1 and Class 2, indicating that messages only attracting a small audience might be very different from the messages which receive huge numbers of retweets. After adding "Temporal" features, shown in the third column, the performance on Class 0 still holds but improvements in Class 1 and Class 2 are observed with some decrease of performance of Class 3. One explanation is that unlike popular messages which receive tens of thousands of retweets, normal messages only attract a small audience and users lose interest in them very quickly. Therefore, temporal features have a stronger effect on these messages with low and medium volume of retweets, compared to highly popular messages. User activity features can further improve the performance marginally, shown in the fourth column in Table 2.

Acknowledgments

This material is based in part upon work supported by the NSF under Grant Number IIS-0545875.

4. REFERENCES

[1] E. Khabiri, C.-F. Hsu, and J. Caverlee. Analyzing and predicting community preference of socially generated metadata: A case study on comments in the digg community. In *ICWSM*, 2009.

[2] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *SocialCom*, 2010.