Prediction of Retweet Cascade Size over Time

Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev,
Pavel Serdyukov, Gleb Gusev, Andrey Kustarev
Yandex
Leo Tolstoy st. 16, Moscow, Russia
{kupavskiy, ostroumova-la, umnov, kaathewise, pavser, gleb57,
kustarev}@yandex-team.ru

ABSTRACT

Retweet cascades play an essential role in information diffusion in Twitter. Popular tweets reflect the current trends in Twitter, while Twitter itself is one of the most important online media. Thus, understanding the reasons why a tweet becomes popular is of great interest for sociologists, marketers and social media researches. What is even more important is the possibility to make a prognosis of a tweet's future popularity. Besides the scientific significance of such possibility, this sort of prediction has lots of practical applications such as breaking news detection, viral marketing etc. In this paper we try to forecast how many retweets a given tweet will gain during a fixed time period. We train an algorithm that predicts the number of retweets during time T since the initial moment. In addition to a standard set of features we utilize several new ones. One of the most important features is the flow of the cascade. Another one is PageRank on the retweet graph, which can be considered as the measure of influence of users.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Twitter, Retweet

Keywords

Retweet cascade, information diffusion, influence

1. INTRODUCTION

Twitter [1] is a very popular online microblogging service, where users share information via tweets. Tweet is a short message that can contain text or link to some external resource. Most of the users receive information through the news feed. If you want to see one's news in the feed, you start *following* him. Thus there is a structure in organization of Tweeter accounts called the *follower graph*. In addition to just reading the news feed users can forward tweets by retweeting them. Retweeting plays significant role in the spread of information [14].

In this paper we study *retweet cascades*. Retweet cascades serve as a valuable data source for the study of information diffusion in Twitter and word-of-mouth diffusion in general. One of the possible goals of such a study is to predict the retweet cascade size in the future based on the information we gain from the past. We address this problem, i.e., we try to predict the number of retweets the tweet will gain during the time T since the initial tweet. Besides the importance of such prediction for sociology, this sort of prognosis has lots of practical applications.

Namely, we consider two tasks of predicting the cascade size at the moment T. The first task allows of utilizing the information available at the initial moment only. The second one enables to consider several first cascade nodes already "infected" at some moment T_0. In both cases we use the information about users' activity during the training period. Intuitively, the second prediction should be more accurate. Both problems are novel to the best of our knowledge. To solve both problems we propose several features in addition to standard ones.

One of the possible motivations behind such study is viral marketing. Suppose you spread an advertisement and you want to get a sufficient volume (e.g. 1000 retweets) within a day. Then you choose the set of initial users and you can try to predict (the first problem), whether you get 1000 retweets or not. Moreover, if you wait for some time and use the initial spread of the cascade (the second problem), then you can make the prediction more accurate.

For this purpose we train an algorithm which utilizes both social (that depend on the user that posted the tweet) and text features of the tweet. While the content features are essentially the same as in [7] and [9], several new social features are considered.

First important new feature is the PageRank of the user in the retweet graph. Retweet graph consists of users as nodes, and two users A and B are connected by a (directed) edge A→B if user A retweeted some tweet of B. Second one is the flow of the retweet cascade, which is described in Section 4. Several others are described in Section 5.

The contribution of this paper is four-fold. First, we put forth two novel prediction tasks: prediction of cascade size over time T since the moment of the initial tweet and analogous prediction task, but utilizing the information about the

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

cascade growth up to the moment T_0. Second, we propose several new features for solving these prediction tasks and train a machine learning algorithm that shows good performance. Third, as one of these features, we propose a novel measure of user's influence: PageRank in the retweet graph. Fourth, we propose an epidemiological model for retweet cascades growth.

The remainder of the paper is organized as follows. We review the related work in Section 2. In Section 3 we describe the collected data. In Section 4 we give the definition of a retweet cascade and of the flow of a cascade. In Section 5 we describe the set of features and the prediction model we propose. In Section 6 we present experimental results.

2. RELATED WORK

Different properties of Twitter follower graph and of retweet cascades are described in [8]. Bakshy et al. [3] studied the individual influence and tweet spread in Twitter. They found that the best prediction of the size of future retweet cascades that user generates is based on the average size of cascades of her tweets in the past and that manually extracted content features do not improve the quality of popularity prediction of the tweet. We focus on the prediction of tweet's popularity, although all of our new factors can be used to measure individual influence. Unlike the work [3], we analyze the content of the tweet automatically, similar to how it was done by Petrovic et al. [9] and Naveed et al. [7]. As authors of [3] claim, the features they used are relatively poor predictors of the cascade size in the future. To improve the quality of such prediction we make the problem more contextual and utilize the information about the initial spread of the cascade, not only about the initial seed.

Prediction of the fact that the tweet will be retweeted at least once was done in [9] and [7]. In our work we do not consider tweets that did not gain any retweet, instead, we try to distinguish between the tweets of different "non-zero" popularity. The set of features we use includes those from these two works and from [3]. In addition, we consider several new features: PageRank in the retweet graph, the flow of the cascade and some time-sensitive and cascade spread sensitive features.

Hong et al. [6] predicted the popularity of tweets based on both social and content features. They classified the tweets into four categories with the following number of retweets: $0, < 100, [100, 9999], \geq 10000$. The set of features they used seems to be rich but unclear since they do not describe the whole set explicitly. However, they achieved good results only for smallest and largest categories. We predict the exact number of retweets the tweet will gain and then classify the tweets into 10 categories according to the number of retweets.

Comparing with works [3], [6], our classification task is more general since we predict the number of retweets the tweet will gain during a certain time period. The use of novel features and the information about the initial spread of the tweet helps to make the classification more accurate.

Kwak et al. [8] and Cha et al. [4] analyzed different user rankings and found that rankings of users based on the number of followers and on the number of retweeted messages differ greatly. Different other rankings of users were proposed: PageRank based on the follower graph [8]; TwitterRank [13], which is a topic-sensitive PageRank based on follower graph with weighted edges; ranking based on how fast the information flows to the account from a random user [11]. However, in several articles (e.g., by Bakshy et al. [3]) it was shown that the best feature for prediction of the future retweetability of the tweet is the average retweet ratio of the initial user. Based on this, we propose a new measure of influence: PageRank of a user in retweet graph, which utilizes both the information of the average retweet ratio and the underlying influence network. However, the comparison of this and other measures is beyond the scope of this paper.

Retweet cascade is a particular case of information diffusion. Different models were suggested for modeling information diffusion. Song et al. [11] modeled information flow using continuous-time Markov chain. Asur et al. [2] modeled the growth of the number of tweets in a trending topic using stochastic multiplicative process. In [10], [12] authors analyzed information diffusion from epidemiological point of view. Steeg et al. [12] investigated the reasons why social epidemics do not spread to a significant fraction of users. To define the flow of the cascade, which is one of the features that we use, we also implicitly use some sort of epidemiological model. As opposite to [12], we consider susceptibility depending on the time that passed since the tweet appeared in the user's feed. Goyal et al. [5] also considered epidemiological model with susceptibility that depends on time, but their approach was different. They predict who will retweet whom rather than the size of the cascade, and they forecast a single retweet if the influence at a given moment is greater than a certain threshold. The comparison of our and other epidemiological models for the tweet spread is of interest but it is also beyond the scope of this paper.

3. DATA

For our study we use two data sources. First, we collected all public tweets over two month period March 1 2012 – April 30 2012 using data from the Twitter "firehose", the complete stream of all tweets. Out of the data for the first six weeks, we extracted all the ordered pairs of users who did a retweet via "retweet" button during these six weeks, and the time of each retweet. We call this dataset D_1, which we use to obtain the information about users' retweet activity. The dataset D_1 contains 750M pairs of users and 1.5B retweets.

Second, out of the last two weeks of the two-month period we obtained a stratified sample (see [3]) of 2000 tweets in English. That is, we put all the tweets that had at least one retweet into 10 logarithmic bins according to the size of their retweet tree. Then we extracted 200 tweets out of each bin. Such a stratification ensures that our sample would reflect the full distribution of the cascade sizes. Then we collected all the information about the corresponding cascades. Namely, we downloaded the times of the retweets, identifiers of corresponding participant users and for each participant we extracted the list of his followers. This dataset D_2 contains 1.3M participating users and 135M followers. We use it to train and to test the machine learning algorithm.

4. RETWEET CASCADES

There are two different ways to make a retweet in Twitter. The first one is via Twitter "retweet" button. The second one is to copy the original message and write "RT @username" or "via @username" etc. We examine only retweets of the first type, since in this case the data is more reliable. For this type of retweets Twitter shows only the first retweet in the feed. That is, given an initial tweet of an arbitrary user,

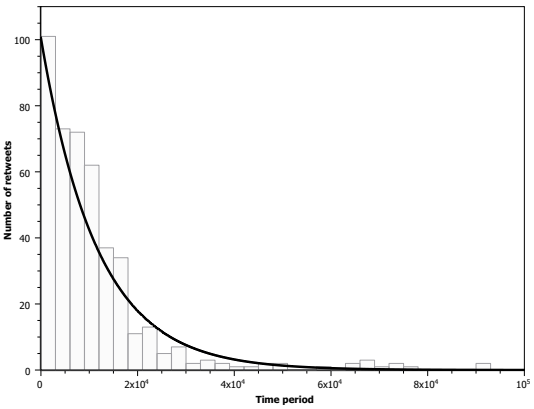

Figure 1: The distribution of time delay of retweet for a sample user

the feed of another user A shows only the earliest retweet among all the retweets of this tweet made by the followees of the user A.

Retweet cascade of a tweet M is the graph, whose vertices are users that retweeted the tweet via "retweet" button. For two users A, B participating in the cascade, we draw the edge $A \to B$ if B follows A and A was the first user among the followees of B who retweeted M. That is, edges go in the opposite direction compared with the follower graph. According to this definition, retweet cascade is a tree, so we also call it *retweet tree*.

In the subsequent paragraphs we define the flow of a cascade which serves as an important feature for our prediction task. For this definition we use a sort of epidemiological model which is of interest in itself. First, for each follower of participating users we define the function of his activity and then the activity of edges between this follower and his followee participating the cascade. We assume that the activity for different incoming edges of one user differs only by a constant. The total activity of an edge during certain time period can be interpreted as the probability to be infected through this edge during this time period. Informally, the flow $F_s(S_0, S)$ of the initial part of a cascade as it was at the moment s is the sum of activities from the moment S_0 to the moment S over all edges between users who were infected before s and their followers, so it is an approximation of the expected number of retweets the tweet will gain from the moment S_0 to the moment S.

Next, we give the formal definition. Consider the initial part of some retweet cascade as it was at moment s. Suppose it consists of users $A_1, \ldots, A_m$, who retweeted the tweet at times $t_1 < t_2 \ldots, < t_m < s$ respectively. For each user A_i, consider the set $\mathcal{B}_i$ of his followers. Since Twitter shows only the first retweet in the feed, for each user we determine the set of followers for whom this user was the first to retweet the tweet. Formally, we construct sets $\mathcal{B}'_i = \mathcal{B}_i \setminus \cup_{j=1}^{i-1} \mathcal{B}_j = \{B_i^1, \ldots\}$. For each user B_i^j we assign a function $f_{B_i^j}(t)$ of the user's activity. To do this, we consider all retweets done by this user during the 6-week period (see Section 3) and the time differences between the initial tweet and its retweet. Typically the distribution of the time delay is approximately exponential (see Figure 1). We model the empirical distribution for the user B_i^j by exponential distribution with parameter λ_{ij}. The parameter is chosen according to the maximum likelihood, that is, $\frac{1}{\lambda_{ij}} = \frac{1}{l}\sum_{i=1}^{l} t_i$, where l is the total number of retweets done by the user and t_i is the time delay for the i-th retweet. We put $f_{B_i^j}(t) = \lambda_{ij} e^{-\lambda_{ij} t}$.

Next, we assign a function $g_e(t)$ to each edge $e : B_i^j \to A_i$ in the follower graph. We have $g_e(t) = c_e f_{B_i^j}(t)$, where

$$c_e = \frac{\sharp(\text{tweets of } A_i \text{ that were retweeted by } B_i^j)}{\sharp(\text{tweets of } A_i)}.$$

Finally, we define the *flow* $F_s(S_0, S)$ of the cascade as it was at the moment s (here $S \geq S_0 \geq s$):

$$F_s(S_0, S) = \int_{S_0}^{S} \sum_{i=1}^{m} \sum_{j} g_{B_i^j \to A_i}(x - t_i) \mathrm{d}x. \qquad (1)$$

5. PREDICTING RETWEET CASCADE SIZE OVER TIME

We want to predict the size of the cascade of the tweet at the moment T given the initial spread of the cascade up to the moment T_0. For this purpose we learn a gradient boosted decision tree model. This model approximates the logarithm of the size of the cascade at the moment T, minimizing mean square error on training examples.

We use four groups of features.

Social features of the initial node (S): number of followers, friends, favorites, number of times the user was listed, is the user verified, number of posts, the date of account creation, average global and local retweet ratio (from the dataset D_1), weighted and not weighted PageRank in the retweet graph. The local retweet ratio is the average number of retweets per tweet done by the user's followers. To calculate the last two features in this group we consider the retweet graph (based on the dataset D_1). For not weighted PageRank we assign equal weights to each edge $A \to B$, for the weighted PageRank we assign weights proportional to the number of tweets of B retweeted by A. All of these features except for two PageRanks were used in [3], [7] or [9].

Content features (C): length of the tweet, number of mentions, hashtags, URLs, positive and negative terms, positive and negative smileys, exclamation and question marks, valence, arousal, dominance. The last three factors are based on Affective Norms of English Words (ANEW) dictionary. These features are standard (for example, see [7, 9]).

The features from the following two groups are used for such tasks for the first time. We fix some small time period t' to measure the delta of characteristics during this short period.

Time-sensitive features of the initial node (TS): average global and local retweet ratios up to the moment t' and T after a tweet posted by the user, the flows of the cascade $F_0(0, t'), F_0(0, T)$ (see Equation (1)).

Denote by CT_0 the set of users already infected at the moment T_0 (including the initial node). We also denote by T'_0 the time the last user from CT_0 was infected.

Features of the infected nodes up to the moment T_o (I): the number of infected users at the moment T_0 ($|CT_0|$), sum of average retweet ratios of the users from CT_0, sum of average retweet ratios up to the moment t' and T of users from CT_0, sum of PageRank over users from CT_0, the average and the total number of followers of users from CT_0, the flows $F_{T'_0}(T_0, T), F_{T'_0}(T_0, T_0 + t'), F_{T'_0}(T'_0, T'_0 + t')$ (see Equation (1)).

	Baseline	+ new features
$T_0 = 0,\ T = 4$ m	0.865	0.728
$T_0 = 0,\ T = 15$ m	0.981	0.957
$T_0 = 0,\ T = 1$ h	1.059	1.038
$T_0 = 0,\ T = 1$ w	1.243	1.226
$T_0 = 15$s, $T = 4$ m	0.865	0.647
$T_0 = 15$s, $T = 15$ m	0.981	0.796
$T_0 = 15$s, $T = 1$ h	1.059	0.919
$T_0 = 15$s, $T = 1$ w	1.243	0.838
$T_0 = 30$s, $T = 4$ m	0.865	0.601
$T_0 = 30$s, $T = 15$ m	0.981	0.488
$T_0 = 30$s, $T = 1$ h	1.059	0.638
$T_0 = 30$s, $T = 1$ w	1.243	0.838

Table 1: mean square error of the logarithm of the predicted cascade size at moment T

6. EXPERIMENTAL SETUP

We execute the prediction for $T_0 = 0, 15, 30$ seconds and for $T = 4, 15$ minutes, one hour and one week. We fix t' equal to 30 seconds. We run a gradient boosted decision tree model with 2000 iterations. We do 5-fold cross-validation. First, we approximate the natural logarithm of the size of the cascade at moment T, minimizing mean square error. We take factors from C and S except for PageRanks as our baseline. The results are shown in Table 1. Note that the case $T_0 = 0$ is the case when we do not use any information about the initial spread of the tweet, i.e., we do not use the set of features I.

One can see that for the prediction from moment $T_0 = 0$ (first task) features from the set TS and PR-features from the set S give an improvement comparing with the baseline, and the improvement is greater for smaller T (for $T = 4$m the improvement is around 17%). What is more interesting, the information about the initial spread of the cascade (even about 15 seconds) helps to make the prediction 25-30% more precise.

Next, we evaluate the contribution of the flow features and PageRank-based features. First, we consider the case $T_0 = 0, T = 1$h and train the algorithm using all features (except for the set I) and using all but one. If we exclude unweighed PageRank or the flow $F_0(0, t')$ then the error is approximately the same as for the whole set of features ($\approx$ 1.038). If we exclude $F_0(0, T)$ or weighted PageRank, then the error becomes slightly bigger ($\approx$ 1.041). Note that the error for baseline factors is 1.059. Second, we consider the case $T_0 = 0, T = 4$m, in which the new features gave most noticeable improvement over the baseline. If we exclude any of the two flows or the two PageRanks then the error becomes much bigger: $\approx$ 0.773 instead of 0.728. We can conclude that these features are more important for a short-term prediction.

Third, we want to find the quality of prediction of the large cascades. We fix $T_0 = 30$s and $T = 1$h. We train two binary classification algorithms that sort out tweets that gain more than 4000 retweets and $[1600, 3999]$ retweets during this time. The number of retweets corresponds to the two bins from our stratified sample that contain tweets with the largest number of retweets. The results for F_1 score are shown in Table 2. The columns correspond to different runs of the algorithm: utilizing only baseline features, all features, all features except for flow features, all features except for PageRank features. It is easy to see that the quality of the prediction with all features is much better than with baseline features and that flow and PageRank features make a substantial contribution.

Tweet class	Baseline	all	no flow	no PR
$[1600, 3999]$	0.659	0.775	0.76	0.761
≥ 4000	0.436	0.67	0.657	0.632

Table 2: F_1-score

7. CONCLUSION

In this work we study the spread of the retweet cascades. We bring forward the new task of predicting the size of the cascade at moment T since the initial tweet. We consider two variations of this task, where in one of them we utilize the information about the initial spread of the cascade. We show that this information helps to improve the quality of the prediction substantially. The prediction is done using machine learning techniques. We propose several new factors that improve the quality of the prediction even when we do not use the information about the initial spread. One of the most important new factors is the PageRank of the user in the retweet graph. We suggest this factor as a measure of user's influence. Another important new factor is the flow of the cascade. To define it we use an epidemiological model for retweet cascade spread.

In the future we plan to study the cascade growth more in detail. Moreover, we plan to compare different measures of influence with the one we propose in this paper.

8. REFERENCES

[1] http://twitter.com

[2] Sitaram Asur, Bernardo A. Huberman, Gabor Szabo, Chunyan Wang *Trends in Social Media : Persistence and Decay*, ICWSM 11

[3] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, Duncan J. Watts *Identifying 'Influencers' on Twitter*, WSDM 11

[4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi *Measuring User Influence in Twitter: The Million Follower Fallacy*, ICWSM 11

[5] Amit Goyal, Francesco Bonchi, Laks V.S. Lakshmanan *Learning Influence Probabilities In Social Networks*, WSDM 10

[6] Liangjie Hong, Ovidiu Dan, Brian D. Davison *Predicting Popular Messages in Twitter*, WWW 11

[7] Nasir Naveed, Thomas Gottron, Jerome Kunegis, Arifah Che Alhadi, *Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter*, WebSci'11

[8] H Kwak, C Lee, H Park, S Moon, *What is Twitter, a Social Network or a News Media?*, WWW 10

[9] Sasa Petrovic, Miles Osborne, Victor Lavrenko, *RT to Win! Predicting Message Propagation in Twitter*, ICWSM 11

[10] Daniel M. Romero, Brendan Meeder, Jon Kleinberg *Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex contagion on twitter*, WWW 11

[11] Xiaodan Song, Yun Chi, Koji Hino, Belle L. Tseng *Information Flow Modeling based on Diffusion Rate for Prediction and Ranking*, WWW 07

[12] Greg Ver Steeg, Rumi Ghosh, Kristina Lerman *What Stops Social Epidemics?*, ICWSM 11

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He *TwitterRank: Finding topic-sensitive influential twitterers*, WSDM 10

[14] Shaomei Wu, Jake M. Hofman, Winter A. Mason, Duncan J. Watts, *Who Says What to Whom on Twitter*, WWW 11