Predicting the Audience Size of a Tweet

Andrey Kupavskii and **Alexey Umnov** and **Gleb Gusev** and **Pavel Serdyukov**
Yandex
Leo Tolstoy st. 16, Moscow, Russia
{kupavskiy, umnov, gleb57, pavser}@yandex-team.ru

Abstract

We study information dissemination in Twitter. We present an analysis of two important characteristics of so called " retweet cascades": retweet count and "show" count, i.e., number of users that receive the tweet in their feed. We show that these two measures behave differently. We describe three models that are aimed to predict the audience size of a tweet: first one utilizes only the data available at the moment of the initial tweet, the second one utilizes the spread of the cascade up to some moment while the third one is an "online prediction".

Introduction

Twitter [1] is a very popular online microblogging service, which serves more often as a source of information rather than as a social network (Kwak et al. 2010). One important feature in Twitter is that each user has its own personal news feed. If user A wants to read all tweets of user B, he starts *following* B. After that all further tweets of B are translated into the news feed of A. Each tweet has the "retweet" button, which allows users to add the tweet to their own tweet stream (with the original authorship properly attributed), so that people who follow the user can read it as well, causing the so called *retweet cascades*. There are several different quantities reflecting the popularity of a tweet. One is the number of retweets the tweet received. Another one is the number of *shows*, or the tweet's *audience size* – the number of users who got the opportunity to see the tweet in their news feed.

Among the first two measures, the first one received significant amount of attention ((Bakshy et al. 2011), (Hong, Dan, and Davison 2011), etc.), while the second one was not considered previously to the best of our knowledge. However, in different scenarios both of these two measures may be the right measure of popularity. For example, for a viral video it seems more important to receive as many retweets as possible, while for a new brand that launches an advertising campaign it may be more important to receive as many shows of the tweets with its name as possible, in order to increase the brand awareness. In our work we study the interconnection between these two measures and show that indeed they differ strongly.

Our main task is to predict the future influence of a tweet in terms of the audience size of the tweet. For this purpose we train an algorithm, which utilizes social (that depend on the user that posted the tweet), text and initial spread features of the tweet. We briefly describe a possible application of such prediction to marketing. Suppose a company spreads an advertisement and wants to get a sufficient volume of activity (e.g. 1000 retweets or 100,000 shows) within a day. Based on the prediction of future popularity one could estimate the cost of the campaign and select a set of initial spreaders. Moreover, if one waits for some time and uses the information about the initial spread of the cascade, then one could conclude whether the campaign is running well or additional posts are required. We also briefly discuss possible methods for computationally efficient approximation of the number of shows of a tweet at the early stages of its dissemination.

The contribution of this paper is three-fold. First, we compare two different measures of tweet influence: number of retweets and number of shows. Second, we introduce several novel prediction tasks: prediction of the number of shows over time T since the moment of the initial tweet, analogous prediction task, but utilizing the information about the cascade growth up to the moment T_0, and online prediction. Third, we propose several new features, train a machine learning algorithm for solving these prediction tasks and analyze the importance of different features for the prediction.

The remainder of the paper is organized as follows. We review the related work in the next section. In the third section we describe the collected data. In the fourth section we compare the two measures of tweet popularity. In the fifth section we describe the prediction model we propose. In the sixth section we present experimental results of the prediction. In the last section we conclude this work and outline the research directions to follow in the future.

Related work

Different properties of Twitter follower graph and of retweet cascades are described in (Kwak et al. 2010). Study of properties of retweet cascades in detail is done in (Lerman, Ghosh, and Surachawala).

Bakshy et al. (Bakshy et al. 2011) studied the individual influence and tweet spread in Twitter. They found that the

[1] http://twitter.com

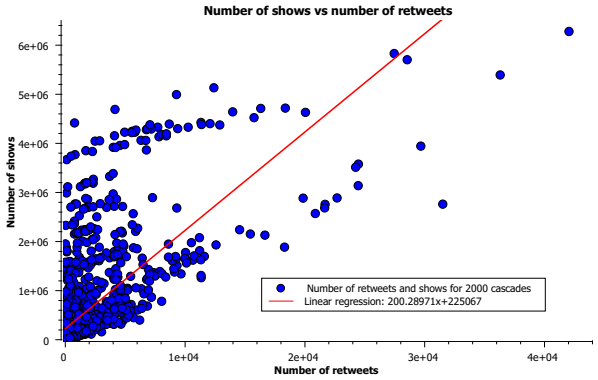

Figure 1: Number of retweets and number of shows for 2000 epidemics

best prediction of the size of 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 (Bakshy et al. 2011), we analyze the content of the tweet automatically, similar to how it was done by (Petrovic, Osborne, and Lavrenko 2011). As (Bakshy et al. 2011) 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 (Petrovic, Osborne, and Lavrenko 2011). Hong et al. (Hong, Dan, and Davison 2011) classified tweets into four categories according to the number of retweets they received. In our work we try to distinguish between the tweets of different "non-zero" popularity. We focus on the prediction of the exact number of retweets or shows rather than on the classification. The set of features we use includes those from these two works and from (Bakshy et al. 2011). In addition, we consider several new features: PageRank in the retweet graph and some time-sensitive and cascade spread sensitive features.

In comparison to (Bakshy et al. 2011), (Hong, Dan, and Davison 2011), we focus on the novel task of prediction of the number of shows, i.e., the tweet's audience, rather than the number of retweets. Our regression task is more general since we predict the number of retweets and shows 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.

In (Kupavskii et al. 2012) authors predicted the exact number of retweets the tweet receives. They introduced the group of initial spread features. In this work, we focus on the prediction of the number of shows. We use several new features and study a novel task of "online" prediction of a tweet's audience size. Szabo and Huberman (Szabo and Huberman 2010) studied the activity of users and popularity of news stories on Digg (http://digg.com). They found out that the early popularity of news linearly correlate with the overall popularity.

In (Kwak et al. 2010) and (Cha et al. 2011) authors 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. We show that the analogous statement is true for the popularity of tweets, that is, that the number of retweets and shows that a tweet received behave in a different way.

Data

For our study we use two data sources. First, we collected a sample of 12B public tweets over two month period March 1 2012 – April 30 2012 using data from the Twitter API. We used the data from the first six weeks to calculate features discussed in the fifth section. Namely, 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 and used it to obtain the information about users' retweet activity. In total we have 750M pairs of users and 1.5B retweets.

Out of the last two weeks of the two-month period we obtained a stratified sample (see (Bakshy et al. 2011)) of 2000 tweets in English. That is, we put all the tweets that were written during the first week (out of these two) and had at least one retweet during these two weeks into 10 logarithmic bins according to the size of their retweet tree (as it was on April 30). Then we extracted 200 tweets out of each bin. Such a stratification procedure ensures that our sample would reflect the full distribution of the non-zero cascade sizes. Then we collected all the information about the corresponding cascades: the times of the retweets, identifiers of corresponding participant users and the list of followers for each participant. It contains 1.3M users that made retweet and 135M unique followers. We use the information about the 2000 cascades to calculate the exact number of shows, to compare the two measures of tweet popularity and also to train and test the machine learning algorithm.

Tweet popularity: Retweets vs. Shows

There are different ways to make a retweet in Twitter, but now the most popular is via Twitter "retweet" button. We examine the retweets of these type. Twitter shows only the first retweet of this kind in the feed of a user. That is, given an initial tweet of an arbitrary 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.

The number of shows the tweet received is the number of users who got the opportunity to see the tweet in their news feeds. At first, we compare the two measures of influence for the tweet. We calculate the coefficient of determination R^2. The coefficient we got that way is equal to 0.45 (in general it could be between 0 and 1), which means that these two measures do not correlate strongly, and so they need different prediction algorithms. Figure 1 shows the dependency between the number of retweets and shows as well as the resulting approximation using linear regression.

Next, we randomly sampled 20 epidemics with large amount of retweets and compared the growth of the number of shows and retweets over time. The resulting figures

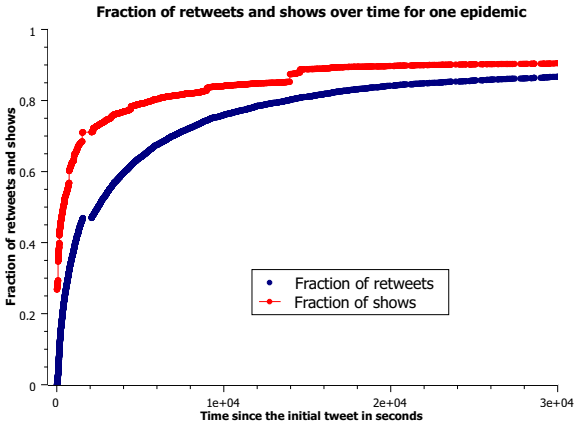

Figure 2: Fraction of retweets and shows received over time for a typical large epidemic

are similar, so on Figure 2 we present this dependency for one typical epidemic. We expect that the prediction of the number of shows should be more precise since at any given time moment we are provided with a larger portion of tweet dissemination measured in terms audience size.

Predicting the number of shows and retweets

We want to predict the popularity of the tweet at the moment T given the initial spread of the cascade up to the moment T_0. Namely, we consider two prediction tasks for each of two measures of tweet popularity we study. These tasks differ in what kind of information about the cascade we use. **The first task** uses only the information available at the tweet share moment. **The second task** uses also the information about the spread of the tweet up to the moment T_0. For each task we learn the gradient boosted decision tree model (Hastie, Tibshirani, and Friedman 2009). We use four groups of features.

Social features of the initial node (S, 12 features): 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 training part of our dataset), show ratio, weighted and unweighted PageRank in the retweet graph (Kupavskii et al. 2012). The local retweet ratio is the average number of retweets per tweet done by the user's followers. Show ratio of the initial user is the average of the number F, where F is the sum of numbers of followers of users that retweeted a given tweet of the initial user. To calculate the last two features in this group we consider the retweet graph (based on the training part of our dataset). For unweighted 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 and show ratio were used in (Bakshy et al. 2011) or (Petrovic, Osborne, and Lavrenko 2011).

Content features (C, 13 features): 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 features are the sums of the corresponding weights of all tweet words that appear in Affective Norms of English Words (ANEW) dictionary. These features were used in (Petrovic, Osborne, and Lavrenko 2011).

We fix some small time period t' to measure the increment of characteristics during this short period. **Time-sensitive features of the initial node (TS, 6 features)**: average global and local retweet ratios up to the moment t' and T after a tweet posted by the user, show ratio up to the moment t' and T.

Features of the infected nodes up to the moment T_0 (I, 12 features): $|CT_0|$, where CT_0 is the set of users already infected at the moment T_0 (including the initial node); sums of local and global retweet ratios of the users from CT_0, sum of average show ratios of the users from CT_0, sum of weighted and unweighted PageRanks over users from CT_0, the number of shows at the moment T_0, show and retweet transmissibility and the scale of the tweet at the moment T_0. The transmissibility of the tweet is the ratio of the tweet popularity (in terms of retweets or shows) at the moment T_0 and average tweet popularity of the initial user at the moment T_0. The scale of the tweet is the transmissibility of the tweet multiplied by the average retweet or show ratio of the initial user up to moment T. That is, this is the expected popularity of the tweet given its spread up to T_0.

Approximating the number of shows

It is computationally complex to calculate the exact number of shows the tweet received up to moment T even if you know the information about the retweets of the tweet since for that we need to store the lists of followers and intersect them online. But we need to calculate the features quickly. So we propose several different approximation algorithms of the number of shows and compare them with each other.

The first approximation is the maximum number of followers among the participants of the epidemic. The second one is the sum of the numbers of followers among the participants. The third one is the "uniform random intersection" approximation. Consider a simple example. Suppose, the epidemic consists of two users, A and B, with the number of followers a and b respectively. We conjecture that the sets of followers of A and B are chosen uniformly and independently from all sets of Twitter users of size a and b respectively. Then the expected intersection of these two sets is equal to ab/M, where $M = 100$ million is the approximate number of active users in Twitter. So the algorithm will estimate the number of shows for this epidemic as $a+b-ab/M$. The fourth one is a random sampling algorithm, which is taken from (Becchetti et al. 2006).

We compare the mean square root error for the approximation of the natural logarithm of the number of shows (for the motivation see the next section) by these methods. The random sampling has mean error 3.32, the maximum — 1.32, the sum — 0.83 and the uniform random intersection — 0.81. The difference between the last two is not large, because the random intersection of two sets of followers is smaller than the real intersection (due to clustering, non-uniform distribution of the number of followees etc.).

Next we put forward the **third prediction task**, which could be called "online prediction". In this case we utilize the information about the initial spread of the tweet except for the exact number of shows the tweet received up to mo-

	retweet	rt user	show	sh user	sh text
0, 15m	0.954	0.956	0.84	0.836	2.826
0, 1 w	1.22	1.218	1.056	1.047	2.78

Table 1: First prediction task

ment T_0, and instead of it use an approximate number of shows. The following features will be used instead of the exact show count in the "online prediction".

Approximate shows count up to the moment $\mathbf{T_o}$ (AS): sum of followers of the users from CT_0, maximum number of followers among users from CT_0, number of shows via "uniform random intersection" and random sampling.

Experiments

We execute the prediction for $T_0 = 0$ and 15 seconds and for $T = 15$ minutes and one week. We fix t' equal to 30 seconds. We run a gradient boosted decision tree model with 200 iterations. We do 10-fold cross-validation. We approximate the natural logarithm of the number of retweets and shows the tweet receives by the moment T by minimizing mean square error. We approximate the logarithm since we want to determine the order of epidemic size rather than the exact size. Moreover, in this case, the algorithm is not biased towards the prediction of the size of large epidemics.The results are shown in Tables 1, 2. In tables we omit T, T_0 and just write their values. If the mean error is equal to $x > 0$, then, roughly speaking, the actual number of retweets (shows) N and the predicted value N' on average have the following relation: $e^{-x}N' \leq N \leq e^{x}N'$. If, e.g., the error is between 0.7 and 1.1, the predicted popularity differs from the actual one in between two and three times.

In Table 1 we present the results for the first task. The first and the third column correspond to prediction that utilizes all features from sets S, C, TS. Columns 2,4,5 correspond to predictions of number of retweets and shows that utilize either only user features (S, TS) or text features (C). One can see that the prediction of the number of shows without content features is even a bit more accurate, which confirms the findings of (Bakshy et al. 2011), while prediction that utilizes only content features gives relatively poor results (however, the prediction is still reasonable).

In Table 2 we present the results for the second and the third tasks. One can see that, first, the prediction becomes more precise if we utilize the information about the initial spread (compare results from Table 1 and Table 2). Second, the quality of the online prediction for the number of shows is a bit better. Third, the precision of the prediction of the number of shows is higher than that for the number of retweets (see both Table 1 and 2), confirming the intuition from the fourth section.

We also present the quality of "online" prediction of the number of shows the tweet got during $T =$ one week using different approximation strategies of the number of shows up to moment $T_0 = 15$ seconds: "no approximation" strategy (1.024), which means that we do not use any of the four features for approximating the number of shows; only "random sampling" strategy (1.023), only "maximum" strategy (1.015), only "sum" strategy (1.023), only "random intersection" strategy (1.02). We see that "maximum" and "random intersection" strategies give the best results.

	retweets	shows	online shows
15s, 15m	0.87	0.827	0.816
15s, 1w	1.14	1.022	1.014

Table 2: Second and the third prediction tasks

Conclusion

In this work we analyze two different measures of tweet popularity: the number of retweets and shows. We explain why both of these measures are of independent interest and find out that these two measures do not have a strong correlation.

We bring forward new tasks of predicting the number of shows the tweet will receive (the size of its audience) at moment T since the initial tweet. We put forth three variations of the prediction: one that utilizes information available at the tweet share moment, the other utilizes the information about the initial spread of the tweet up to some moment T_0 and the third one uses only the features that are easily computable. We provide an analysis of the prediction and the importance of different groups of features. We show that the audience prediction is more precise that the prediction of the number of retweets, which makes it a more reliable measure of tweet popularity. We also study the problem of approximate calculation of the number of shows the tweet received.

In the future we plan to study another measure of tweet influence that is suitable for tweets containing URLs. It is the number of clicks the link receives and corresponds to the size of the audience that not only had the opportunity to see the tweet, or retweeted it, but which actually got engaged with its content.

References

Bakshy, E.; Hofman, J.; Mason, W.; and Watts, D. 2011. Identifying 'influencers' on twitter. In *WWW'11*.

Becchetti, L.; Castillo, C.; Donato, D.; Leonardi, S.; and Baeza-Yates, R. 2006. Using rank propagation and probabilistic counting for link based spam detection. In *WebKDD'06*.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2011. Measuring user influence in twitter: The million follower fallacy. In *ICWSM'11*.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning*. Springer.

Hong, L.; Dan, O.; and Davison, B. 2011. Predicting popular messages in twitter. In *WWW'11*.

Kupavskii, A.; Ostroumova, L.; Umnov, A.; Usachev, S.; Serdyukov, P.; Gusev, G.; and Kustarev, A. 2012. Prediction of retweet cascade size over time. In *CIKM'12*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *WWW'10*.

Lerman, K.; Ghosh, R.; and Surachawala, T. Social contagion: An empirical study of information spread on digg and twitter follower graphs. arXiv:1202.3162.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! predicting message propagation in twitter. In *ICWSM'11*.

Szabo, G., and Huberman, B. 2010. Predicting the popularity of online content. *Communications of the ACM* 53(8):80–88.