Effective and Effortless Features for Popularity Prediction in Microblogging Network

Shuai Gao
gao_shuai@mail.sdu.edu.cn

Jun Ma
majun@sdu.edu.cn

Zhumin Chen
chenzhumin@sdu.edu.cn

School of Computer Science and Technology, Shandong University, Jinan, 250101, China

ABSTRACT

Predicting popularity of online contents is of remarkable practical value in various business and administrative applications. Existing studies mainly focus on finding the most effective features for prediction. However, some effective features, such as structural features which are extracted from the underlying user network, are hard to access. In this paper, we aim to identify features that are both effective and effortless (easy to obtain or compute). Experiments on Sina Weibo show the effectiveness and effortlessness of the temporal features and satisfying prediction performance can be obtained based on only the temporal features of first 10 retweets.

Categories and Subject Descriptors

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

Keywords

Popularity Prediction; Microblogging; Social Network; Classification; Temporal Features

1. INTRODUCTION

The study of popularity prediction on social networks has drawn much attention recently because of its remarkable practical value in a variety of business and administrative applications. It is challenging since there are numerous factors to be considered. Recently, several pioneering work have been made and a wide spectrum of features have been investigated [1, 2]. The recent studies mainly focus on finding the most effective features, while the efforts needed to access these features are neglected. However, some effective features, such as structural features extracted from the underlying user network, are sometimes hard to access. For social networks such as Twitter and Sina Weibo, due to a rate limit on API requests, it is not possible to collect the whole user network. Moreover, it has been shown that the linked structures of social networks do not reveal actual interactions among people [3]. Due to the scarcity of attention, people tend to interact with those few that matter and reciprocate their attention.

In this work, we aim to identify effective and effortless features in popularity prediction problem. By effortless, we mean the features are easy to obtain or compute. The study is conducted on the Sina Weibo, a Twitter-like microblogging network in China. Since most tweets stop spreading in one month, we use the number of "retweets" that a tweet gets in the first month since posted to measure its popularity. Unlike previous work which predict the tweet popularity on hourly or daily basis, we consider an alternate form of popularity prediction problem: Given a tweet and the first few retweets following it, how well can we predict its popularity. We treat this problem as a classification task and train a binary classifier to predict whether a tweet will be popular in the future. By investigating a wide spectrum of features including temporal features derived from retweets chain and two groups of structural features extracted from user network, we show that we can get a satisfying prediction performance by only using the temporal features of first 10 retweets. Note that temporal features can be effortlessly extracted only based on the first few retweets, without need of the knowledge of user network. This finding provides new insights for administrative applications such as media control. Quick decisions can be made effortlessly based on observation of the early stages of diffusion process.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577312.

2. PROBLEM & FEATURES

Problem Definition. Here, we define the popularity of a tweet t, Φ^t, to be the number of retweets that t will get in the first month since posted. Given a tweet t and its first k retweets, our task is to predict Φ^t. Note that predicting the exact value of Φ^t is extremely hard and is often not necessary, we define a popularity threshold ϕ and relax the problem to be a binary classification problem that predicts whether or not Φ^t will exceed ϕ.

Features. We list all features in Table 1. To extract structural features, we first construct a global user network $G = (U, E)$ based on the historical mention relationships, where U is nodes set and E is edges set respectively. The authority scores of users are computed on G using PageRank algorithm. For a given tweet t, we sort all its retweets by ascending time order, forming a chain of retweets. We use u_i^t to denote the author of the i^{th} retweet. Specially, we denote the author of t as u_0^t. Considering the first k retweets in the chain, we denote the union of tweet author and retweet authors as rU_t, i.e. $rU_t = \cup_{i=0}^{k}\{u_i^t\}$. By extracting relationships from G, we form a retweet network $rG_t = (rU_t, rE_t)$. Further, by considering only the retweet authors, $r_sU_t = \cup_{i=1}^{k}\{u_i^t\}$, we construct a strict retweet network $r_sG_t = (r_sU_t, r_sE_t)$. From rG_t and r_sG_t, we extract 10 structural features to characterize the retweet network. The depth of network is the longest length of the path from the tweet author to any retweet author in the network. The reciprocity defined as the portion of co-links is used to measure the tie-strength between users in the network. Based on G, border users bU_t are followers of rU_t

Table 1: Features

Type	Description
Retweet Network Features	(1) density of rG_t, (2) depth of rG_t, (3) reciprocity of rG_t, (4) clustering coefficient of rG_t, (5) number of connected component in r_sG_t, (6) number of connected component (size $>$ 2) in r_sG_t, (7) maximum size of connected component in r_sG_t, (8) authority of u_0^t, (9) average of authority of authors in rG_t, (10) maximum authority of authors in rG_t
Border Network Features	(1) number of border users $\|bU_t\|$, (2) density of bG_t, (3) reciprocity of bG_t, (4) average authority of users in bG_t, (5) maximum authority of users in bG_t, (6) 15-dimension exposure probability vector
Temporal Features	(1) k-dimension time vector: time taken for first k retweet to arrive, (2) max time interval

Table 2: The performance of the prediction task

Methods	Pre.	Rec.	F1	ACC	AUC
Random	0.5000	0.5000	0.4994	0.5000	0.5000
Pos.-% bias	0.5000	0.5000	0.5000	0.5023	0.5000
All features	0.7418	0.7354	0.7364	0.7403	0.7354
- RN Feature	0.7385	0.7321	0.7231	0.7370	0.7321
- BN Feature	0.7330	0.7259	0.7267	0.7311	0.7259
- T Feature	0.6312	0.6285	0.6285	0.6333	0.6285

who still have not retweeted tweet t, i.e. $bU_t = \{u_q | \exists e(u_p, u_q) \in E, u_p \in rU_t, u_q \notin rU_t\}$. By extracting relationships between rU_t and bU_t, we construct a border network $bG_t = (rU_t, bU_t, bE_t)$, which is a bipartite network. From bG_t, we extract 6 structural features to characterize the border network. In exposure probability vector, $P(x)$ $(1 \leq x \leq 15)$ is the ratio of border users who have x edges from users in rU_t to the total number of border users. For temporal features, similar to that in [1], we first preprocess the time for each tweet to mitigate the dependence of the tweet popularity on the time of day when it was posted. We measure the time taken for the first k retweets to arrive and the max time interval between two adjacent retweets in the first k retweets.

3. EXPERIMENTS

Experimental Setting. We use Sina Weibo dataset published by WISE 2012 Challenge. The global user network is constructed based on mention relationships from Jan to Aug 2011 and consists of 10.8 million users and 87.1 million edges. We selected a subset of tweets that receive at least 10 retweets in July 2011. This gave us a dataset of 51,835 original tweets and 4,645,067 retweets from 1,031,899 users. We reserved 50% of the tweets for evaluation, using the other 50% for feature exploration. In our experiments, by setting $k = 10$ and $\phi = 50$, we created an approximately balanced binary prediction problem: given the first 10 retweets of a tweet, will the tweet receive at least 50 retweets finally. We use bagged decision trees with 60 trees as the classifier. We use all features described in Table 1 to train our classifier and compare the results to two baselines. *Random* baseline chooses a tweet's label randomly with no bias and *Positive-percentage bias* baseline chooses a tweet's label randomly with bias equals to the percentage of positive class in the test set (46.63%).

Results. We report classification Precision, Recall, F_1 score, ACC (accuracy) and AUC (area under RUC curve) for each method in the upper part of Table 2. Obviously, combining all features yields the best performance. Further, we perform a stepwise forward feature selection algorithm to identify effective features for prediction and show the top-5 selected features in Table 3. We can see that a relative small set of features can achieve comparable performance to all features. It is reasonable that the best single feature is the maximum authority of authors in rG_t since if the tweet is posted or retweeted by a higher authority user, it will have more chance to be seen and retweeted. Note that, two of the top-5 features are from the temporal feature group and there is a significant performance gain after adding these features. The other two features are from border network feature group and can be interpreted as the number of exposed users and the tie strength between users who have retweeted the tweet and those who have not. Seeing that the top features are from different feature groups, we further check the effectiveness of each feature group by removing each feature group and examining how the prediction performance is affected. The results are shown in lower part of Table 2. We can see that the performance drops significantly when temporal features are removed. On the contrary, when retweet network features or border network features are removed, the performance slightly changes. That indicates the temporal features contribute greatly to the overall performance and the combination of other features is unable to make up the loss. We highlight this by comparing the prediction performance when using "all features" "temporal features" and "without temporal features" of the first k retweets and show the results in Fig. 1. Obviously, the best performance is always achieved by combining all the features. Also, we can see that when the temporal features are removed, the performance of the other features slightly changes when k varies. One explanation is that, for a large portion of tweets which gain at least 10 retweets, maximum authority of authors in rG_t is equal to the authority of u_0^t, which is unchanged when k varies. It is worth noting that the performance gap between all features and temporal features gradually narrows with the increasing of k. When $k = 10$, the performance gap has been reduced to 0.0235. That indicates, by only using the temporal features of the first 10 retweets, we can get a satisfying prediction performance. Extracting structural features needs to first construct the underlying user network, which is sometimes hard to get as mentioned before. However, temporal features can be effortlessly extracted only based on the first few retweets.

Table 3: Results of stepwise forward feature selection. Each row represents the performance for all features listed in that row and above.

Feature added	AUC
maximum authority of authors in rG_t	0.6212
+ time vector [10]	0.6721
+ number of border users $\|bU_t\|$	0.7065
+ time vector [7]	0.7204
+ reciprocity of bG_t	0.7268

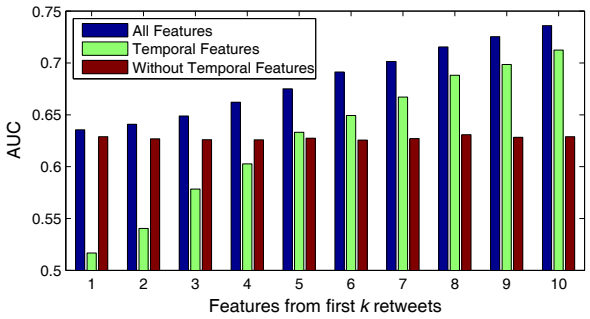

Figure 1: Performance when predicting using only the features derived from the first k retweets.

References

[1] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, August 2010.

[2] Peng Bao, Hua-Wei Shen, Junming Huang, and Xue-Qi Cheng. Popularity prediction in microblogging network: a case study on sina weibo. In *WWW companion*, pages 177–178, 2013.

[3] Bernardo Huberman, Daniel Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *SSRN 1313405*, 2008.