Retweet Behavior Prediction in Twitter

Article · April 2015

DOI: 10.1109/ISCID.2014.187

Feisheng Yang
University of California, San Diego
33 PUBLICATIONS **321** CITATIONS

Retweet Behavior Prediction in Twitter

Dongxu Huang, Jing Zhou
School of Automation
Northwestern Polytechnical University
Xi'an, China
e-mail: huangdongxu21@mail.nwpu.edu.cn
zhoujingnwpu@gmail.com

Dejun Mu, Feisheng Yang
School of Automation
Northwestern Polytechnical University
Xi'an China
e-mail: {mudejun, yangfeisheng}@nwpu.edu.cn

***Abstract*—Retweet, as a main way to spread information in twitter, has been researched in a number of works. Recently research focuses on analyzing the factors of retweet behavior. However, the prediction on retweet behavior is a new challenge which is not well studied in the past. A basic fact is that different people are interested in different kinds of tweets, and they will retweet tweets which they are interested in. First, we collect tweets of different categories from valid account of famous news media as learning corpus. Second, in order to discover user interests, we classify user tweets into different categories by Bayes model. Finally, we measure user interests on tweets of different categories, and predict retweet behavior by interest measurement. This paper extends the previous study on retweet behavior, and we predict user retweet behavior as well as infer user interests. Experiment shows Bayes model has good performance on classifying tweets, and our algorithm achieves more precision than others.**

Keywords-retweet; retweet bahavior prediction; twitter; Bayes model; user interests

I. Introduction

Millions of users publish a huge amount of tweets in twitter daily, and retweeting is deemed as a key mechanism of information diffusion. The study on retweet behavior will help us to understand how the information spread and whom the information will arrive at. A number of previous works analyse the factors of retweet behavior [1-3]. However, a few of works attempt to predict it [4-12]. The main challenges of predict retweet behavior are summarized below. First, the prediction on retweet behavior by tweet text similarity is not accuracy because the content of tweet is limit to 140 characters, even some tweets contains some nonsense words, personal names, and urls. Second, it's difficult to discover user's interests from tweets, so we don't know whether the user likes the special tweet and he will retweet it or not.

We propose to predict retweet behavior by user interests which has not been researched as far as we know. The basic idea is illustrated in Fig 1. Nodes t_{u1}-t_{u5} in the top represent user tweets, nodes c_1-c_3 in the center represents tweet category, nodes t_1-t_4 in the bottom represent other tweets. User tweets are classified into two categories (t_{u1}, t_{u2}, t_{u5}, c_1) and (t_{u3}, t_{u4}, c_2), in which the user interests can be denoted by c_1, c_2, respectively. While other tweets are divided into three categories (t_1, c_2), (t_2, c_1), (t_3, t_4, c_3). In our work, we predict that user maybe retweet t_1 and t_2, and are almost impossible to retweet t_3 and t_4. And the exact process will be discussed in Section 4.

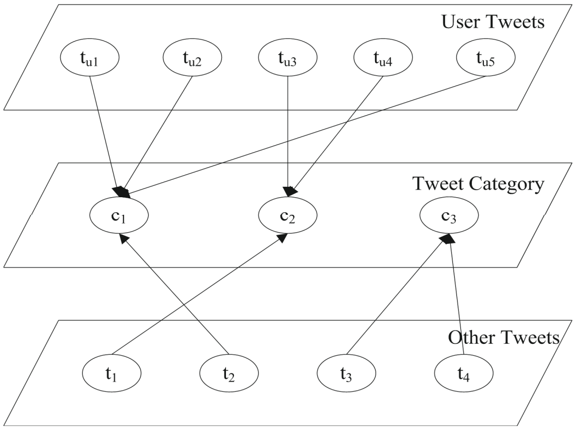

Figure 1. Simple introduction of our method

The contributions of this paper are three folds: (1) we propose predict retweet behavior by user interest. People retweet the tweet because they are interested in it. (2) We infer user interest by classifying user tweets to different categories according to Bayes model. (3) Experimental has been carefully designed to evaluate the effectiveness of user interests discovery methodology and retweet behavior prediction.

The rest of paper is organized as follows: Section 2 provides an overview of state-of-the-art research on retweet behavior. Section 3 defines the problem formally. Retweet behavior prediction method is presented is Section 4, followed by experimental evaluation and analysis in Section 5. Our work and possible future directions are summarized in Section 6.

II. Related Work

Twitter is a popular social network and generates over 340 million tweets everyday. With the availability of large scale Twitter data, we summarize relevant works in this section.

Some researchers analyse the factors of retweet behavior[1-3], Xu select 22 features which can be divided four classes (social-based, content-based, tweet-based, author-based), and apply three classification algorithm to the training model and predict retweet behavior[1]. B. Suh analyses 9 factors which will affect retweet behavior according to carefully experiment and find that URLs as well

978-1-4799-7005-6/14 $31.00 © 2014 IEEE
DOI 10.1109/ISCID.2014.187

as hashtags have strong relationships with retweetability[2]. However, it's complex to consider so many features, and their experiments also show that some features is more important than others, so our work focus on the content of tweet and makes in-depth study on it.

Moving away from retweet factor analysis, researchers have also study on retweet prediction [4-9]. Luo consider a wide range of features and use SVM$^{\text{Rank}}$ to train their retrieval model and predict retweeters[4]. AMAV (Autoregressive moving average model) has been used in [7] and some machine learning algorithm is applied in [8-9]. However, we predict retweet behavior by user interests, and this method has not been well studied before.

Other researchers not only predict retweet behavior, but also explain why they retweet [10-12]. A factor graph model has been used to predict retweet behavior and analyse how the retweet behavior is influenced by factors [10]. Macskassy classifies tweets to different categories by the knowledge of Wikipedia, and builds information propagation model based on tweet categories to study retweet behavior [11]. However, there are so many categories in Wikipedia and it's difficult to merge sub-categories. Then, we classify tweets to 8 categories according to news categories of famous new media, and apply their tweets as learn corpus.

User interests discovery also be studied in [13-14]. However, we infer user interests by classifying user tweets to different categories which is different from their works.

III. Problem Definition

We first introduce the following notations to be used in the rest of paper. $U = \{u_1, u_2, ...u_n\}$ represents the set of twitter users, $C = \{C_1, C_2, ...C_p\}$ represents the set of tweet categories, $T = \{t_1, t_2, ...t_m\}$ is the tweet set, the set of words in t_i is denoted by $T_i = \{w_1, w_2, ...w_l\}$.

Table 1 NOTATIONS

$P(w_j \mid C_i)$	the probability of w_j in C_i
$P(C_j \mid t_i)$	the probability of t_i belongs to C_j
$C(t_i)$	the final category which t_i belongs to
$P(C_j \mid w_k)$	the probability of w_k belongs to C_j
$P(w_k)$	the probability of w_k in all categories of tweets
$P(\text{C}_i)$	the percentage of tweets belongs to C_i
WC_i	the word set of C_i
$F(u, C_l)$	the interest threshold of u in C_l

Table 1 summarizes the notations. The problem is then defined as follows:

Input: Given tweet category dataset G, a user u and all of his tweets, a tweet t from his friends.
Output: Whether u will retweet t.

IV. Retweet Behavior Prediction

A. Tweet Category Feature Description

In order to classify tweets to proper categories, we carefully collect 146875 tweets which belong to 8 categories from 5 famous news media's tweets. We will explain the principle of data collection in Section 5.A.

We apply a matrix M to describe the features of different tweet categories. The columns are the words, and the rows are the tweet categories, and m_{ij} represents the probability of the word w_j in category C_i,

$$m_{ij} = P(w_j \mid C_i) = \frac{N_i(w_j) + \delta}{\sum_{w_j \in WC_i} N_i(w_j) + \delta |WC_i|} \quad (1)$$

Where δ is a smooth parameter, and it is set to 0.5 in our experiment. $N_i(w_j)$ represents the number of word w_j in category C_i.

B. User Tweet Classification

We infer user interests by classifying user tweets into different categories which has been defined before. User tweets are classified into some special categories, and he will like these categories of tweets. We classify user tweets by computing the probability of the tweets belongs to each category as follows.

$$P(C_j \mid t_i) = \prod_{w_k \in T_i} P(C_j \mid w_k) \quad (2)$$

$$P(C_j \mid w_k) = \frac{P(C_j)P(w_k \mid C_j)}{P(w_k)} = P(C_j)\frac{P(w_k \mid C_j)}{\sum_{C_i \in C} P(C_i)P(w_k \mid C_i)} \quad (3)$$

We suppose that the probability of each tweet category is same, and set $P(C_i) = 1$ for convenience. The biggest probability tweet category has been chosen as the final category which the tweet belongs to. We evaluate the classification model in Section 5.B.

C. User Retweet Behavior Prediction

We classify user tweets into different categories and discover user interests in Section 4.2. However, in order to predict user retweet behavior, we have to know how much the user like different categories of tweets and the probability of the given tweet belongs to its final category. Then, first, suppose $P(C_j \mid t_i)$ denotes the probability of t_i belongs to C_j, we calculate the mean of the summation of $P(C_j \mid t_i)\ (C(t_i) = C_j)$, which is the measurement of user interest on tweet category C_j. Second, we compute the probability of the given tweet belongs to its final category. Finally, if the probability greater than the measurement of user interest on the tweet final category, the user will retweet the tweet, else will not.

The algorithm is described as follows in detail:

```
**Input**: tweet category dataset $G$, a user $u$ and all of his tweets { $t_1^u$ , $t_2^u$ , $t_3^u$ , …, $t_n^u$ }, a tweet $t$ from his friends.
**Output**: Whether $u$ will retweet $t$ .
    Step 1: Compute tweet category feature matrix M.
    Step 2: for $i = 1:1:n$
        $P(C_j \mid t_i^u) = \prod_{w_k \in T_i} P(C_j \mid w_k)(1 \le j \le p)$
        $P(C_q \mid t_i^u) = \max(P(C_j \mid t_i^u))\ (1 \le j \le p)$
        $C(t_i^u) = C_q$
    Step 3: for $l = 1:1:p$
        $T_l^u = \{t_r^u \mid C(t_r^u) = C_l\}$
        $F(u, C_l) = \lambda_{C(l)}$
        $\lambda_{C(l)} = \sum_{t_r^u \in T_l^u} P(C_l \mid t_r^u) / |T_l^u|$
    Step 4. If $P(\mathrm{C}(t) \mid t) \ge \lambda_{C(t)}$ Then $u$ will retweet $t$
            If $P(\mathrm{C}(t) \mid t) < \lambda_{C(t)}$ Then $u$ won't retweet $t$
```

V. EXPERIMENTs

In this section, we introduce the datasets which used in the experiment and two baseline algorithms. And then we evaluate the Bayes model for tweet classification. Finally, we apply our methodology and two baseline algorithms on twitter dataset.

A. Dataset and Baseline Algorithms

1) Tweet Category Dataset. This dataset is used for get features of different tweet categories. We choose five famous new media (CNN, USA Today, Times, New York Times, Yahoo News). These famous new media have multiple tweet accounts in twitter, and each accounts publish a category of tweets. We collect eight common categories of tweets from their different tweet accounts, the eight categories are technology, politics, life, sports, entertainment, health, travel, finance.

2) User Tweet Dataset. This dataset is used for evaluating retweet behavior predict algorithm. We collect large amount of users tweets and retweets. However, the users which have few tweets are inactive users and unrepresentative. So we only select 10000 users which publish 500 tweets at least, and use their tweets and retweets as user tweet dataset.

3) Baseline algorithms. We use two baseline algorithms.

The first is TSA (Text Similarity Algorithm). Text similarity used for retweet behavior analysis and prediction has been researched in many works [1, 2, 4, 9, 14]. In our work, we apply TF-IDF (Term Frequency-Inverse Document Frequency) to get user interest vector from all of user tweets, and get tweet vector by lucene[1]. Then we compute cosine similarity between user interest vector and tweet vector, we predict that user will retweet the tweet if the cosine similarity greater than a threshold. The threshold is the average of cosine similarity between user interest vector and themselves tweets.

The second is CTA (Co-Terms Algorithm). URLs and hashtags are the important factors of retweet behavior [1, 2, 9]. The authors like the same website if their tweets contain same URLs. They talk about the same topics if their tweets contain same hashtags. So in CTA, user will retweet the tweets which have same URLs or hashtags with himself.

B. User Tweet Classification Evaluation

The performance of user tweet classification affects the precision of retweet prediction. In order to evaluate user tweet classification, we extract 3000, 6000, 9000, 12000, 15000 tweets from eight categories of tweets in the tweet category dataset respectively as learning corpuses. Then we compute the feature vector of each categories of tweets. We use the tweet category dataset as test data and apply tweet classification algorithm on it. The experiment result on different categories of tweets and different learning corpuses are shown in Fig. 2.

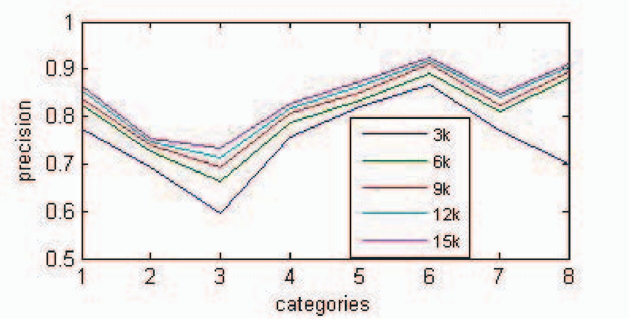

Figure 2. User tweet classification evaluation

Experiments shows the more tweets used to training in our model, the more precision of tweet classification. Then we use 15000 tweets as learning corpus of each category of tweets in the retweet prediction experiments.

C. User Retweet Behavior Evaluation

First, we compute the interest thresholds of 10000 users on eight categories of tweets and then plot the thresholds which has been randomly selected from 100 users in Fig. 3(a), 3(b). The mean and variance of 10000 users thresholds are shown in Table 2.

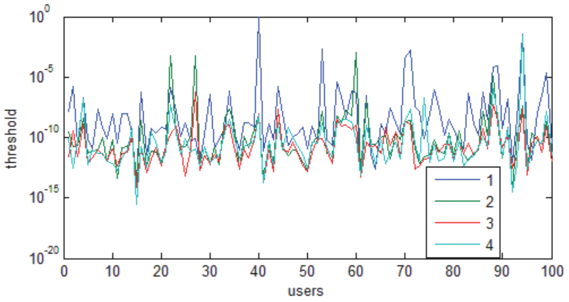

Figure 3(a). Threshold of 100 users on four categories of tweets

[1] http://lucene.apache.org/

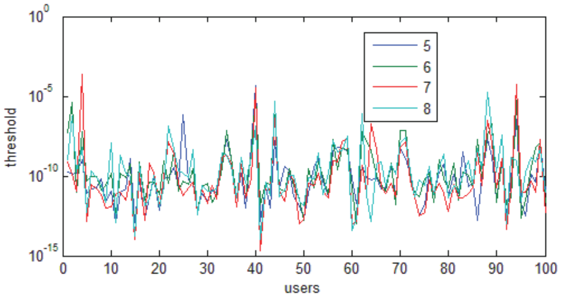

Figure 3(b). Threshold of 100 users on another four categories of tweets

TABLE 2 THE MEAN AND VARIANCE OF THRESHOLDS

	1	2	3	4	5	6	7	8
mean(10^{-4})	29	11	6.88	20	12	21	9.38	14
variance(10^{-4})	2.7539	2.3619	1.0745	5.0090	3.5399	3.5376	2.6794	2.6069

Second, we compute precision of three algorithms in user tweet dataset. We randomly select 100 users result and plot it in Fig. 4. The mean and variance of all users prediction result in three algorithms are shown in Table 3. Experiments result show that UIM is more precise and stable than TSA and CIA.

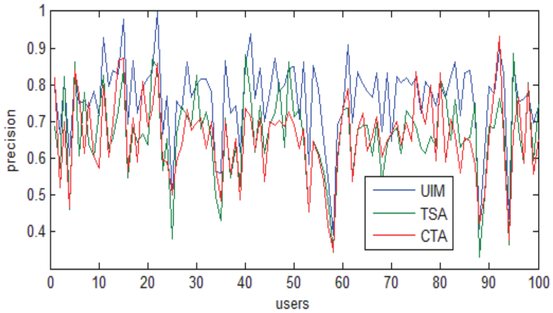

Figure 4. Precision of three algorithms

TABLE 3 THE MEAN AND VARIANCE OF PREDICTION RESULTS

	UIM	TSA	CIA
mean	0.7704	0.7105	0.6102
variance	0.0118	0.0334	0.0534

VI. CONCLUSION

Understanding retweet behavior could help us to know how information spreads in twitter. We propose a novel algorithm that predicts user retweet behavior by user interest. First, we discover the categories of tweets which the user likes. Second, we measure users' interest on the categories which they like. Finally, we compute the probability of the given tweet belonging to special tweet category. If the probability greater than the user interest on the special category, we predict the user will retweet the tweet, else will not. Experiments show our algorithm outperforms other algorithms in precision and stability. The further research will focus on inferring user interests by unsupervised learning.

ACKNOWLEDGEMENT: The work is supported by NSFC NO. 61403311.

REFERENCES

[1] Z. Xu, and Q. Yang, "Analyzing user retweet behavior on twitter", in Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, pp.46-50, ASONAM, 2012.

[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 Social Computing, 2010 IEEE second International Conference on, pp. 177-184, IEEE, 2010.

[3] D. M. Romero, B. Meeder and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter", in Proceedings of the 20th International Conference on World Wide Web, pp. 695-704, 2011.

[4] Z. Luo, M. Osborne, J. Tang, and T. Wang, 'Who will retweet me? Finding retweeters in twitter', in Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 869-72, ACM, 2013.

[5] T.-T. Kuo, R. Yan, Y.-Y. Huang, P.H. Kung and S.D. Lin, "Unsupervised link prediction using aggregative statistics on heterogeneous social networks", in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 775-783, ACM, 2013..

[6] S. Petrovic, M. Osborne and V. Lavrenko, "Rt to win! Predicting message propagation in twitter", in Proceedings of the 2011 International AAAI Conference on Weblogs and Social Media, pp.586-589, ACM, 2011.

[7] Z. Luo, Y. Wang and X. Wu, "Predicting retweeting behavior based on autoregressive moving average model", in Web Information Systems Engineering-wise, pp.777-782, Springer, 2012.

[8] L. Zhang, Z. Zhang and P. Jin, "Classification-based prediction on the retweet actions over microblog dataset", in proceedings of the 13th international conference on web information systems engineering, pp. 771-776, 2012.

[9] F. Abel, Q. Gao, G.-J. Houben and K. Tao, "Analyzing user modeling on twitter for personalized news recommendations", in user modeling, adaption and personalization, pp. 1-12, Springer, 2011.

[10] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, and et al, 'Understanding retweeting behaviors in social networks', in Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1633-36, ACM, 2010.

[11] S. A. Macskassy, M. Michelson, 'Why do people retweet? Anti-Homophily wins the day!', in Proceedings of the 2011 International AAAI Conference on Weblogs and Social Media, pp.209-216, ACM, 2011.

[12] F. Pezzoni, J. An, A. Passarella, J. Crowcroft and M. Conti, "Why do i retweet it? An information propagation model for microblogs", Social Informatics, pp. 360-369, Springer, 2013.

[13] J. Wang, W. X. Zhao, Y. He and X. Li, "Infer user interests via link structure regularization", ACM Transaction on Intelligent System and Technology, 5(2), pp.1-22, 2014.

[14] B. Wang, C. Wang, J. Bu, C. Chen, W. V. Zhang, et al, "Whom to mention: Expand the diffusion of tweets by@ recommendation on micro-blogging systems", in proceedings of the 22nd international conference on World Wide Web, pp. 1331-1340, International World Wide Web Conferences Steering Committee, 2013.

[15] J. Weng, E.-P. Lim, J. Jiang and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers", in Proceedings of the third ACM International Conference on Web Search and Data Mining, pp. 261-270, ACM, 2010.