Who Will Retweet Me? Finding Retweeters in Twitter

Zhunchen Luo[†], Miles Osborne[‡], Jintao Tang[†] and Ting Wang[†]
[†]College of Computer, National University of Defense Technology
410073, Changsha, Hunan, CHINA
[‡]School of Informatics, The University of Edinburgh
EH8 9AB, Edinburgh, UK
zhunchenluo@nudt.edu.cn, miles@inf.ed.ac.uk, {tangjintao, tingwang}@nudt.edu.cn

ABSTRACT

An important aspect of communication in Twitter (and other Social Networks) is message propagation – people creating posts for others to share. Although there has been work on modelling how tweets in Twitter are propagated (retweeted), an untackled problem has been **who** will retweet a message. Here we consider the task of finding who will retweet a message posted on Twitter. Within a learning-to-rank framework, we explore a wide range of features, such as retweet history, followers status, followers active time and followers interests. We find that followers who retweeted or mentioned the author's tweets frequently before and have common interests are more likely to be retweeters.

Categories and Subject Descriptors

Information Systems [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Retweeter, Twitter, Propagation Analysis

1. INTRODUCTION

Twitter is a popular microblogging service which attracts over 500 million registered users and generates over 340 million tweets daily[1]. People not only consume information but also publish or share opinions or stories themselves. An interesting mechanism in Twitter is *retweeting* – re-posting someone else's tweet. Most studies in retweeting concentrate on predicting whether a tweet will be retweeted [5, 13, 1] or understanding retweeting behavior [11, 4, 18]. However, we are still largely ignorant about **who** will retweet a post. We call such people *retweeters*. Here, we consider the problem of finding retweeters in Twitter.

Modelling who will retweet a post deepens our understanding of how information flows in Social Networks. These factors can be a complex interaction between the message content itself, the original author and the person who propagates that message. Understanding who will share a post is of interest to advertisers and media companies. This is clearly a hard task as much of the information necessary for prediction is hidden from us, or changes over time.

Here, we treat finding retweeters as a ranking problem which retrieves followers retweeting a certain post. We use a standard machine learning approach to learn a ranking function for followers that uses a range of features. These features include: the retweet history of the original author, the social status of her followers, when they are active and their interests. We demonstrate that our ranking approach using all features can achieve MAP significantly better than a **random** baseline which ranks the followers randomly and a baseline which ranks followers by the number of times she **r**etweeted the author's **p**revious **t**weets before. In particular, we find that the retweet history and the similarity between the content of the tweet and the posting times of followers are most effective for finding retweeters.

2. RELATED WORK

The popularity of Twitter, easy access to data and the unique characteristic of retweeting have made it a hot research domain recently. Related work can be divided into predicting retweets and retweeting behavioral analysis.

2.1 Predicting Retweets

In Twitter, messages deemed important by the community propagates through retweets. There is much work related to predicting whether a tweet in general will be retweeted. Suh *et al.* [16] firstly examined a number of features that might affect retweetability of tweets using a large amount of data. They found that, amongst content features, URLs and hashtags have strong relationships with retweetability. Petrovic *et al.* [13] used a machine learning approach based on the passive-aggressive algorithm to predict whether a tweet would be retweeted in the future. They found that the tweet content, the number of times the author was listed, how many followers they had and whether the author was verified were useful features for this task. Hong *et al.* [5] proposed a method to predict whether a tweet would be retweeted and also estimated the number of times it would be retweeted. Zaman *et al.* [19] used a collaborative filtering approach to

[1] `en.wikipedia.org/wiki/Twitter`

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

predict for a pair of users whether a tweet written by one will be retweeted by the other user. They found that the identity of the source of the tweet and retweeter were the most effective features for predicting future retweets. Naveed *et al.* [12] trained a prediction model to forecast for a given tweet its likelihood of being retweeted based on its contents. They introduced and evaluated a method to determine the 'interestingness' of microblog messages. Their experimental results showed that this interestingness made a message on Twitter worth retweeting. Artzi *et al.* [1] proposed a model for predicting the likelihood of responding which includes retweeting and replying.

2.2 Retweeting Behavioral Analysis

Moving away from predicting retweets, researchers have also studied broad characteristics of retweeting. boyd[2] *et al.* [4] interviewed Twitter users on the reasons why they retweet, and on what they retweet the most. Yang *and* Counts [17] constructed networks based on user name mentions when measuring how the network structure affected information diffusion in Twitter. Yang *et al.* [18] analyzed how retweeting behaviors was influenced by factors such posting time. They found that almost 25.5% of tweets posted by users are actually retweeted from friends who had blogs. Macskassy *and* Michelson [10] studied a set of Twitter users over a period of a month and sought to explain the individual information diffusion behaviors, as represented by retweets. They found that content based propagation models could explain the majority of retweet behaviors they saw in their data. Starbird *et al.* [15] examined microblogging information diffusion activity during the 2011 Egyptian political uprisings. They demonstrated how remote individuals participated in Egypt's 2011 revolution through low-risk, social media- enabled activities. Comarela *et al.* [3] identified factors that influence users' response or retweet probability. They found previous response to the tweeter, the tweeters' sending rate, the freshness of information, the length of tweet could affect users' response.

3. METHOD

Given a tweet t from user u and her followers $Followers(u)$, our goal is to learn a function F that estimates the likelihood of follower f_i ($f_i \in Follower(u)$) retweeting t in future. We treat this as a ranking problem and find the top-k followers who are most likely to retweet a given post. This is because tweets might be retweeted by variable numbers of people and potentially we might want to rank them. Note that this could be treated as a classification problem without loss of generality.

To generate a good function F which ranks $Followers(u)$ according to whether they are likely to be retweeters of a given post, we investigate a wide range of features. We develop features in a learning-to-rank scenario [7]. Learning to rank is a data driven approach which incorporates a set of features in a model for ranking task [8, 9]. Every follower f_i is tagged whether she retweeted t in training data. This gives author-follower pairs. From these author-follower pairs a set of features related to the possibility of being retweeters is extracted.

3.1 Features

When retrieving retweeters, we consider the following feature families:

3.1.1 Retweet History (RH)

Intuitively, some follower f_i who retweeted (mentioned) a user u before is likely to retweet u's tweets again. We develop two features: the number of times f_i previously retweeted u's tweets (called **Num_fRu**) and the number of times f_i mentioned u before (called **Num_fMu**). This captures the idea that two users may talk to each other and so may be more inclined to propagate posts in the future. We also consider the situation when two users mutually repost each other. We count the number of times u retweeted and mentioned f_i's tweets before (called **Num_uRf** and **Num_uMf** respectively) as two other features for our task. Finally, some people (often spammers) only repost other people's information and hardly post original texts. We use the ratio of a follower's tweets which are retweets (or contain mention '@'), called **Ratio_retweet** (**Ratio_mention**), as two features.

3.1.2 Follower Status (FS)

Information propagation is often from higher status[3] users to lower status users [2]. We randomly investigated more than 100,000 retweets and found only 38.8% of retweets are from users posting fewer tweets to users posting more tweets; only 23.8% of retweets are posted from users containing fewer followers to users containing more followers; only 0.04% retweets are from unverified users to verified users. These statistics show that users with different social status have different retweeting behavior. We include this information as follows: the number of tweets user f_i has ever written (called **Posts**), the number of followers user f_i has (called **Followers**), the number of friends user f_i has (called **Friend**), the times user f_i has been listed (called **Listed**) and whether f_i is verified (called **Verified**).

3.1.3 Follower Active Time (FAT)

Twitter users do not often interact with other users (very) late at night. Furthermore, posts made late at night are often not seen by other users the next day due to them being replaced by more recent tweets[4]. To explore this statement, we considered which time users actually replied to other users. After randomly sampling 10,000 tweets that were replies, we found that only 12.4% of users replied to tweets between the hours of 00:00 and 06:00.

We model this information using two features. The first captures when two users are in the same timezone (called **Timezone**). We use the proportion of f_i's tweets posted before which were in the same hour interval to the tweet t's posting time as another feature called **PostTimeConsis**. This feature could capture the consistency of follower's posting time habit to t's posting time.

3.1.4 Follower Interests (FI)

A follower retweeting a tweet might indicate some connection between the two people; for example, they might share common interests ("we both love cats"), be having an off-line relationship etc. This information can be a valuable hint when finding retweeters. We present a similarity

[2] Dana Boyd spells her name in publications with lower case letters. This is not a typo.

[3] Where *status* could mean a celebrity.

[4] For example, Twitter only displays the most recent posts.

model to capture shared interests and represent the user f_i's previous tweets as a simple weighted bag-of-word of terms (based on Tfidf score). We can then compute a cosine angle between the user f_i's previous tweets and the tweet t (based on vector space model [14]). We call this feature the **SimInterest**. When calculating the value of feature **SimInterest**, for each pairs of tweet t and follower f_i's previous tweets, we filtered the top 100 high frequent words and the words which appear less than 5 times in more than 6 millions tweets in our collected data.

4. EXPERIMENTS

4.1 Dataset and Experimental setting

To the best of our knowledge, there is no annotated dataset of retweeters in Twitter. Therefore, we created a new dataset for this task[5]. We randomly selected 500 English tweets which had been retweeted at least once by the respective author followers. Tweet posting times were from September 14th, 2012 to October 1st, 2012. Since retweeting is time sensitive (half of retweeting occur within an hour, and 75% of retweeting under a day [6]) we rechecked these tweets again after one day to investigate which followers retweeted the original tweet. Due to time limitation of the Twitter API and the large number of followers for popular users, it is infeasible for us to collect all follower information for each author. Therefore, we only collected 100 recent new followers from all the followers of tweet authors. We also collected 200 recent tweets of each follower. As previously explained, we assigned a binary label to every follower indicating whether they retweeted a post. Note that some users may not be available due to privacy concerns preventing us from achieving total recall. Importantly, because we do not exhaustively collect all follower information for each user, it is possible that a tweet is retweeted by a user that we do not have information for. This will also reduce overall recall. Summary statistics of the data are listed in Table 1.

Total tweets which have been retweeted	500
Average number of followers per tweet	81.15
Total retweeters	257
Total non-retweeters	40317

Table 1: Retweeters Data Statistics

We use SVM^{Rank} when training our author retrieval model[6]. We use a linear kernel for training and report results for the best setting of parameters. In order to avoid overfitting the data we perform 10-fold cross-validation in our dataset. There are several metrics that can be used to measure the performance of ranking. In this paper, we use *Mean Average Precision* (MAP) as the evaluation metric.

4.2 Results

4.2.1 Baselines

We choose two baselines. The first one simply ranks followers randomly and we call it *Random*. For the second baseline, we note that followers who previously had a history of retweeting might do this in the future. Therefore, we ranks followers by the number of times they **r**etweeted the author's **p**revious **t**weets before (called RPT).

4.2.2 Retrieval Performance

Our ranking methods use different feature families: **Retweet History**, **Follower Status**, **Follower Active Time** and **Follower Interests** (called RH, FS, FAT and FI respectively).

Table 2 shows retrieval performance according to which family of related features we use. The main result is that user similarity (*Follower Interest, FI*) is the most beneficial information source, followed by retweeting history (*RH*) and then social status of followers (*Follower Status, FS*). Information about active time of the follower is not useful for this task. Finally, using all feature families together (called *All*) achieves the best performance, improving MAP by 301.4% over the *Random* and 25.6% over the *PRT*.

Next we consider in detail the effect of individual features when finding retweeters in Twitter. Table 3 shows the results using individual features when ranking[7]. We can see that all the ranking methods based on the features from **Retweet History** feature family yield significant improve this task. We also found the feature **PostTimeConsis** helps. Additionally, the effectiveness of **SimInterest** shows most users retweet another user's posts based on content.

4.2.3 Examples

Here is an example showing the usefulness of user retweet history:

> *We are having a bake sale today in the Student Union from 11-2! Come buy a midday snack from the Pretty Poodles!*

The author of this tweet was retweeted by a follower who retweeted or mentioned the author's tweets 30 times totally in previous tweets. Our *RH model* ranks this follower in first place.

Here is another example showing the usefulness of user interest similarity:

> *Excited to announce our debut London show. Full details here - `http://t.co/P60Wc3Lj`*

The author of this tweet has a follower who retweeted this post and in turn had posted often about music and performance as shown in previous tweets. This follower was ranked higher by the *FI model*.

5. CONCLUSION

Finding retweeters in Twitter can help deliver information to other people more efficiently and effectively. This is a new task and our results will open the way for follow-up research better understanding how Social Media works. In our work we find the historical retweeting records of author and followers, the followers' status and the similarity between the content of the tweet and followers' previous tweets are effective information for this task.

[5] This dataset is available at `https://sourceforge.net/projects/retweeter/`

[6] `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

[7] The MAP results of *PRT* and *Num_fRu* ranking models are different, since we use the weight of feature Num_fRu for *PRT* to rank followers directly and take this as a feature for learning to rank in *Num_fRu* model.

	MAP(%)		MAP(%)
Random	2.17	FAT	2.91
PRT	6.93	FI	8.12^*
RH	6.27^*	All	$8.71^{*\dagger}$
FS	3.66^*		

Table 2: Performance of Ranking Methods for Finding Retweeters based on Different Feature Family. A significant improvement over the *Random* ranking method and *PRT* ranking method with a star (*) and a dagger (†) respectively ($p < 0.05$).

	MAP(%)		MAP(%)
Random	2.17	Posts	3.79^*
PRT	6.93	Followers	2.37
Num_fRu	6.83^*	Friends	2.03
Num_fMu	7.08^*	Listed	2.17
Num_uRf	6.20^*	Verified	2.34
Num_uMf	7.62^*	Timezone	2.37
Retweet_Ratio	4.45^*	PostTimeConsis	2.86^*
Mention_Ratio	3.05^*	SimInterest	8.12^*

Table 3: Performance of Ranking Methods for Finding Retweeters based on Different Feature Alone. A significant improvement over the *Random* ranking method and *PRT* ranking method with a star (*) and a dagger (†) respectively ($p < 0.05$).

In the future our work is using new features to improve the performance of finding retweeters. These may include whether the intimate friends of follower retweet the certain tweet, whether the location information of the tweets affects the followers, etc.

Acknowledgements

The research is supported by the National Natural Science Foundation of China (Grant No. 61170156, 61202337 and 60933005).

6. REFERENCES

[1] Y. Artzi, P. Pantel, and M. Gamon. Predicting responses to microblog posts. In *HLT-NAACL*, pages 602–606, 2012.

[2] M. Cha, A. Mislove, and P. K. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, pages 721–730, 2009.

[3] G. Comarela, M. Crovella, V. Almeida, and F. Benevenuto. Understanding factors that affect response rates in twitter. In *HT*, pages 123–132, New York, NY, USA, 2012. ACM.

[4] danah boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*, pages 1–10, 2010.

[5] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *WWW (Companion Volume)*, pages 57–58, 2011.

[6] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600, New York, NY, USA, 2010. ACM.

[7] T.-Y. Liu. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331, Mar. 2009.

[8] Z. Luo, M. Osborne, S. Petrovic, and T. Wang. Improving twitter retrieval by exploiting structural information. In *AAAI*, 2012.

[9] Z. Luo, M. Osborne, and T. Wang. Opinion retrieval in twitter. In *ICWSM*, 2012.

[10] S. A. Macskassy and M. Michelson. Why do people retweet? anti-homophily wins the day! In *ICWSM*, 2011.

[11] M. Nagarajan, H. Purohit, and A. P. Sheth. A qualitative examination of topical tweet and retweet practices. In *ICWSM*, 2010.

[12] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci*, 2011.

[13] S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. In *ICWSM*, 2011.

[14] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[15] K. Starbird, L. Palen, and L. Palen. (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *CSCW*, pages 7–16, 2012.

[16] 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/PASSAT*, pages 177–184, 2010.

[17] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. In *ICWSM*, 2010.

[18] Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *CIKM*, pages 1633–1636, 2010.

[19] T. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, pages 599–601, 2010.