Real-time Scholarly Retweeting Prediction System

Zhunchen Luo[†*] and **Xiao Liu**[‡*]
[†]Information Research Center of Military Science, PLA Academy of Military Science
100142 Beijing, China
zhunchenluo@gmail.com
[‡]School of Computer Science and Technology, Beijing Institute of Technology
100081 Beijing, China
xiaoliu@bit.edu.cn

Abstract

Twitter has become one of the most import channels to spread latest scholarly information because of its fast information spread speed. How to predict whether a scholarly tweet will be retweeted is a key task in understanding the message propagation within large user communities. Hence, we present the real-time scholarly retweeting prediction system that retrieves scholarly tweets which will be retweeted. First, we filter scholarly tweets from tracking a tweet stream. Then, we extract Tweet Scholar Blocks indicating metadata of papers. At last, we combine scholarly features with the Tweet Scholar Blocks to predict whether a scholarly tweet will be retweeted. Our system outperforms chosen baseline systems. Additionally, our system has the potential to predict scientific impact in real-time.

1 Introduction

The volume of information about scientific papers is enormous on Twitter, and most data is real-time, even before the paper content is published and shortly after the notifications of acceptance. Besides, lots of scholars post tweets to express their excitement when their papers got accepted (Priem and Costello, 2010). We call the tweets that imply accepted papers scholarly tweets (*STs*) and the rest non-scholarly tweets (*NSTs*). Retweeting is an action of reposting others' tweet by using the *retweet* button on Twitter or other mechanism. To help understand the message propagation within large user communities, we develop a real-time scholarly retweeting prediction system.

Our task is to predict whether a *ST* will be retweeted. The problem of retweeting prediction has attracted more and more attention. Zhang et al. (2016) propose a deep learning method to predict retweeting. However, due to the special and structural ways using combinations of different Tweet Scholar Blocks (*TSBs*) encoding scholarly information about papers, venues, and authors, different methods should be explored to solve our task.

In this work, we propose a real-time scholarly retweeting prediction system by exploiting *TSBs* and scholarly features. We only focus on retweets made using the *retweet* button in Twitter. Under this circumstance, the tweet-retweet connection is unambiguously and can be retrieved directly by Twitter's API. At first, we trace a data stream by tracking "paper accepted" in Twitter using the Twitter API, but there are some *NSTs* in the data stream, so we build a classification model to filter *ST tweets*. It is investigated that most *STs* consist of text blocks called *Twitter Scholar Blocks* (*TSBs*) indicating meta data, and we build a sequence tagger to extract *TSBs* to gather metadata information. At last, we build a binary classification model by combining *TSBs* with scholarly information in Twitter to predict whether the *ST* will be retweeted. Experimental results show that our system outperforms chosen baseline systems and has the potential to predict scientific impact in real-time.

---

* indicates equal contribution.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

2 Real-time Scholarly Retweeting Prediction System

2.1 System Overview

Given a tweet t, our goal is first to learn a function STF that estimate the likelihood of whether t is a scholarly tweet, then learn a function RP to estimate the probability of whether t will be retweeted. By incorporating with the *TSBs* and scholarly features, we use the system to predict whether the *STs* will be retweeted. The framework of our approach is shown in Figure 1.

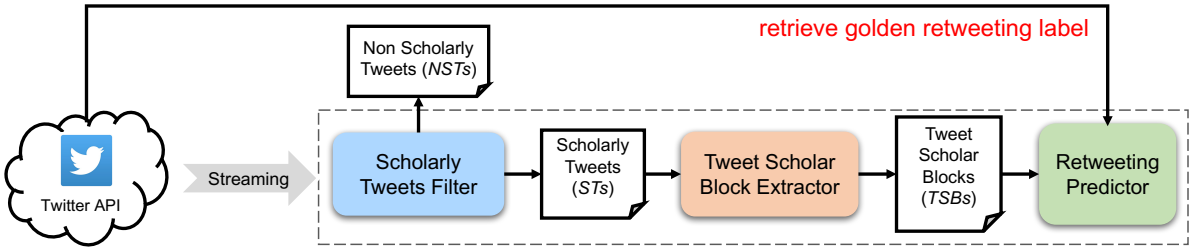

Figure 1: Framework of Our Approach

2.2 Scholarly Tweets Filter

We regard filtering *STs* from the data stream as a classification problem. In our scholarly tweets filter (*STF*) module, we build a classification model based on support vector machine.

To capture the information in social networks, we design a feature **user's scholarly membership of academic institutions** by examining whether user descriptions contain one of the high-frequency words of academic institution names in Wikipedia (we choose top sixty words in experiments). Additionally, we design **bag of words, words with trending symbols and length of the tweet** as features. We also find that almost no one would hide happiness if her paper were accepted, and we use a tweet-specified sentiment analysis API[1] to generate **sentiment labels** for tweets.

2.3 Tweet Scholar Block Extractor

Inspired by previous works on structuring tweets (Luo et al., 2012; Luo et al., 2015), we investigate that researchers post *STs* in structural ways using combinations of different Tweet Scholar Blocks (*TSBs*) encoding scholarly information about papers, venues, and authors. In this work, we propose six types of *TSBs*: **Author**, the names of authors; **Title**, the title of the paper; **Venue**, the short or entire name of the venue; **Time**, the time when the venue will be held; **Place**, the place where the venue will be held; **Other**, the rest part of tweet text. An example of extracted *TSBs* of a tweet is given in Figure 2.

In our tweet scholar block extractor (*TSBE*) module, we build a sequence tagger based on conditional random fields with BIO schema. We use **tokens starting with "@", surrounded by pairwise symbols, capitalized, trending symbols, POS-Tagging labels and NER labels** as our features. Tokens starting with "@" in *STs* are often mentioned co-authors. Besides, the paper titles usually occupy up to 40% text content which is often surrounded by pairwise symbols or all capitalized to show different formats.

2.4 Retweeting Predictor

In our retweeting predictor (*RP*) module, we build a classification model based on support vector machine (*SVM*). Apart from using text information generated from the extracted *TSBs* as our features, we take scholarly features from social networks information in Twitter into account. Apart from extracted *TSBs*, we categorize the rest scholarly features into following two categories:

Author Social Features: Previous work shows that the overall impact of all co-authors should have the potential to influence a paper's quality and popularity (Dong et al., 2015). We use extracted *Author* type of *TSBs* to find the authors in *ST tweets*. We think the influence of an individual is related to her **friend's**

[1] https://dev.exploreyourdata.com/index.html

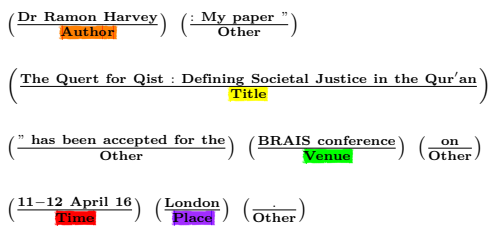

Figure 2: An Example of Extracted Tweet Scholar Blocks

number, followers number, and statuses quantity. To show the influence of a group, we calculate the **sum, maximum value, minimum value and average value** of influences of all participants in that group. In spite of these, we design a binary feature indicating **whether a user is verified**as verification is used by Twitter mostly to confirm the authenticity of celebrity accounts.

Venue Popularity Features: Different venues have large differences in their influences. Since the well-respected venues are better platforms for researchers to publish their work or results, our intuition is that better sites help scholars spread their scientific impact more. Scholars often use Twitter as a note-taking tool (Mapes, 2016) during venues, so **the number of statuses in the venue topic** may reflect the popularity and influence of the site. Considering the developments and the trends of the venues, we also take **the total historical quantity of statuses** into account.

3 Experiments

Predicted	Golden	User	Tweet
not	not	@Neonatal_Brain	New ultrasound marker for #brain growth. Paper is accepted. Nice and easy for both fetal and neonatal measurements. https://t.co/W17Wcnqv4L
yes	yes	@danieldekok	\o/. ACL short paper by me and Erhard Hinrichs on dependency parsing with topological fields (and BiDi LSTMs) accepted.
yes	not	@manaalfar	Short paper on non-distributional word vectors accepted at ACL 2015 #chinacl2015 #acl2015 #nlproc
not	yes	@dimazest	finally I got an email from emnlp and it's positive, the paper got accepted!

Table 1: Examples of predicted scholarly tweets and the golden labels of whether they will be retweeted

3.1 Data Preparation and Experiment Settings

To evaluate our system, we first track a tweets stream posted from Jan. 2012 to Apr. 2018 by tracking the key phrase "paper accepted" using Twitter API. randomly crawl 6,500 tweets Then we randomly sample 6,500 tweets and manually label them as *STs* and *NSTs* for training scholarly tweets filter. Next, we choose 1,400 original *STs* out of them by checking their "retweeted_status" attributes are empty from Twitter API. We use tweet-specific annotators (Owoputi et al., 2013; Ritter et al., 2011) to tokenize those tweets and get pos-tagging and NER labels, then manually label *TSBs* in BIO schema for training our tweet scholar block extractor. Last but not least, we get the golden labels of whether an original *ST* will be retweeted by finding the corresponding retweets. Additionally, five-fold cross-validation is used in our experiments and accuracy is used as the evaluation metric.

3.2 Baseline Comparison and Feature Selection

We choose two baselines, the one is random prediction (*Random*), the other is an CNN model (Zhang et al., 2016) (*SUA-ACNN*). Then we compare the result of using *TSBE* and golden *TSBs* with *RP* (*TSBE+RP* and *Golden+RP* respectively). To find the best feature conjunction, we use a greedy feature selection method in which we first choose the best feature set out of several randomly generated sets and then iteratively append features that yield better performance. The setting of using best feature set is called

TSBE+RP_Best and *Golden+RP_Best* respectively. Results are shown in Table 2. Besides, Table 1 demonstrates some predicted examples of our *TSBE+RP_Best* system.

Approach	Accuracy
Random	62.43%
SUA-ACNN	76.29%
RP	90.36%
TSBE+RP	87.43%
RP_Best	94.50%
TSBE+RP_Best	90.57%

Table 2: Comparing With Baselines and the Best Feature Conjunction

Overall, our system outperforms the baseline, and it is feasible to predict scientific impact in Twitter in real time. Moreover, the performance of *TSBE+RP* is lower than the performance of *RP* on manually labeled TSBs, because the errors produced in *TSBE* might affect the performance of *RP*. Besides, the best feature conjunction consisted of *Sum Friends Count*, *Sum Followers Count*, *Max Followers Count*.

3.3 Ablation Study

To find the effectiveness of each feature and which features are in particular highly valued by *RP_Best*, we also removed each feature from *RP_Best* and *TSBE+RP_Best* respectively to evaluate the effectiveness of each feature by the decrement of accuracy.

By comparing the results shown in Table 3, we can see that *Sum Followers Count* is very effective to our *RP_Best*. The reason might be that *Sum Followers Count* is more suitable to stand for the influence of the authors' group.

Approach	Accuracy
RP_Best	94.50%
RP_Best-Sum Friends Count	89.57%
RP_Best-Sum Followers Count	88.93%
RP_Best-Max Followers Count	89.14%
TSBE+RP_Best	90.57%
TSBE+RP_Best-Sum Friends Count	86.71%
TSBE+RP_Best-Sum Followers Count	85.14%
TSBE+RP_Best-Max Followers Count	86.43%

Table 3: Comparing Results by Decaying Every Feature One by One

4 Conclusion

In this paper, we propose our real-time scholarly retweeting prediction system which solves the scholarly tweets retweeting prediction problem. We introduce the three modules in our system: scholarly tweets filter, tweet scholar block extractor and retweeting predictor. In addition, our system has the potential to predict scientific impact in real-time. Sufficient experimental results demonstrate that our model outperforms the baseline systems. Hope our system can help researchers to stand on the shoulders of right giants.

Acknowledgement

We thank the anonymous reviewers for their helpful comments. We also thank our annotators for giving suggestions when accomplishing the dataset and holding helpful discussions. This work is supported by National Natural Science Foundation of China (No. 61602490).

References

Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. 2015. Will this paper increase your *h*-index?: Scientific impact prediction. In *Proceedings of the WSDM 2015*, pages 149–158.

Zhunchen Luo, Miles Osborne, Sasa Petrovic, and Ting Wang. 2012. Improving twitter retrieval by exploiting structural information. In *Proceedings of the AAAI 2012*, pages 648–654.

Zhunchen Luo, Yang Yu, Miles Osborne, and Ting Wang. 2015. Structuring tweets for improving twitter search. *JASIST*, 66(12):2522–2539.

Kristen Mapes. 2016. A qualitative content analysis of 19, 000 medieval studies conference tweets. In *Proceedings of the SIGDOC 2016*, page 48.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the NAACL-HLT 2013*, pages 380–390.

Jason Priem and Kaitlin Light Costello. 2010. How and why scholars cite on twitter. *Proceedings of The Asist Annual Meeting*, 47(1):1–4.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the EMNLP 2011*, pages 1524–1534.

Qi Zhang, Yeyun Gong, Jindou Wu, Haoran Huang, and Xuanjing Huang. 2016. Retweet prediction with attention-based deep neural network. In *Proceedings of the CIKM 2016*, pages 75–84.