Real-Time Scientific Impact Prediction in Twitter

Zhunchen Luo[1(✉)], Jun Chen[1], and Xiao Liu[2]

[1] Information Research Center of Military Science,
PLA Academy of Military Science, 100142 Beijing, China
zhunchenluo@gamil.com, 13501239808@126.com
[2] School of Computer Science and Technology,
Beijing Institute of Technology, 100081 Beijing, China
xiaoliu@bit.edu.cn

Abstract. As the number of scientific publication is getting larger and larger, scientific impact prediction has become an urgent need. However, traditional scientific impact prediction, which is mainly based on longtime accumulated citation networks, metadata and the whole text of papers, is relatively hysteretic and can hardly fit the rapid development of technology. Moreover, Twitter has become one of the most import channels to spread latest technique information because of its fast information spread speed. The advantage of publishing new messages in real-time can compensate the imperfections of traditional scientific impact prediction methods. Therefore, we propose a new approach to predict scientific impact in Twitter in real time before publishing the paper content. After filtering scholarly tweets (*ST tweets*), and extracting Tweet Scholar Blocks (*TSBs*) indicating metadata of papers to help predict scientific impact in real time, author social features, venue popularity features, and title features are exploited to predict whether the article will increase h-index of its first author after five years. Our model achieves an outstanding result that its best accuracy is 80.95%. The best feature conjunction consists of the sum of friends and followers of all the co-authors, followers count of the first author and title embeddings. And the amount of followers of all the co-authors is the most critical feature. Our finding reveals that Twitter has the potential to predict scientific impact in real time. We hope that real-time scientific impact prediction in Twitter can help researchers to expand their influences and more conveniently "stand on the shoulders of giants".

Keywords: Twitter · Scientific impact prediction · Real-time

1 Introduction

Scientific impact prediction has become an urgent need since the number of scientific publication is getting larger and larger. As an instance, the number

Z. Xu et al. (Eds.): Big Data 2018, CCIS 945, pp. 108–123, 2018.
https://doi.org/10.1007/978-981-13-2922-7_7

of e-print publications in *arXiv*[1] has exceeded 1,354,091. Scientific publications are spread in various channels and platforms, such as Twitter, Mendeley, printed journals.

Twitter is now one of the biggest social networks, and the vast volume of tweets posted on Twitter per day is highly attractive for information retrieval purpose. There not only is a tremendous amount of unrevealed information about scientific papers in Twitter but also are lots of scholars post tweets to express their excitement when their papers got accepted [12,22]. We call the tweets that imply accepted papers scholarly tweets (*ST tweets*) and the rest non-scholarly tweets (*NST tweets*).

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. However, previous work shows that most scientific impact prediction works are based on citation networks [16], metadata of papers [15], or text content of articles [32], and those methods are quite time-consuming, as the analysis requires the publication content of the paper. The wish to predict the scientific impact of a newly published paper may be delayed to a great degree. On the contrary, for example, the *ST tweet*[2] illustrated in Fig. 1 published on May 25, 2016, implies that the paper: "Domain Adaptation for Authorship Attribution: Improved Structural Correspondence Learning" co-authored by Manuel Montes is accepted by the Association for Computational Linguistics 2016 (*ACL 2016*). Notifications of acceptance[3] of long papers were delivered on May 24, 2016. And the conference was held from August 7 to August 12, 2016. Apparently, the *ST tweet* was posted before the date of publication which shows content. If we can predict the scientific impact of the paper once it has been accepted even before it comes to publication, we can use the real-time prediction to boost later information analyzation in much more advanced.

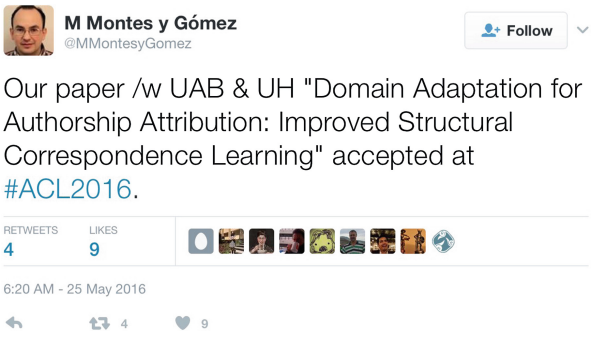

Fig. 1. An example of a *ST Tweet*

[1] https://arxiv.org/.
[2] https://twitter.com/MMontesyGomez/status/735460462758789120?lang=en.
[3] http://acl2016.org/index.php?article_id=9.

Toward this end, we propose a new approach to predict scientific impact in Twitter, so that the impact can be calculated in real time, before the publication of the related paper. At first, we trace a data stream by tracking "paper accepted" in Twitter, but there are some *NST tweets* in the data stream, so we build a classification model to filter *ST tweets*. To predict scientific impact in Twitter in real time, we want to make use of the metadata of papers. It is investigated that most *ST tweets* consist of text blocks called *Twitter Scholar Blocks* (*TSBs*) indicating meta data. According to the investigation, we then build a sequence tagger to extract *TSBs* to gather metadata information. Finally, we build a binary classification model by combining *TSBs* with information in social networks in Twitter to predict whether the paper implied in *ST tweet* will increase the h-index [9] of its first author after five years. The best accuracy of our model is 80.95%, which outperforms the baseline based on citation networks. We find the best feature conjunction consists of the sum of friends and followers of all the co-authors, followers count of the first author and title embeddings, besides, amount of followers count of all the co-authors is the most critical feature.

Contributions: The main contributions of this paper are three-fold:

(1) We show that social networks like Twitter have the potential to predict scientific impact in real time.
(2) We propose *TSBs* in *ST tweets* and a novel approach utilizing *TSBs* to predict scientific impact in real time with 80.95% accuracy.
(3) We discover that the best feature conjunction consists of the sum friends and followers count of all the co-authors, followers count of the first author and title embeddings. And sum followers count of all the co-authors is the most critical feature.

2 Related Work

Our research is related to two aspects of work; the one is traditional scientific impact prediction that is regarded as a regression problem on citation numbers, the other is scientific analysis in social media.

2.1 Regression Scientific Impact Prediction

Scientific impact prediction often seems like a regression problem on citations. Information extracted from citation networks is widely used. [5] use temporal elements and topological elements to predict future citation. [16] use encoding method based on citation network of Scopus database. [26] investigated the factors determining the capability of academic articles to be cited in the future using topological analysis of citation networks.

Text information seems to be popular in recent years. [32] consider predicting measurable responses to scientific articles primarily based on their text content. [15] analyze the usefulness of rich information derived from the full text of the

articles through a variety of approaches, including rhetorical sentence analysis, information extraction, and time-series analysis and they combine metadata and whole text to achieve a better result.

There are also works that combine these two types of information. [31] adapt a discriminative approach that can make use of any text or metadata and show that lexical knowledge offers substantial power in modeling out-of-sample response and forecasting response for future articles. They show that various social factors influence written scientific communication and they can uncover these factors by measuring language similarity between articles.

Although approaches mentioned above perform efficiently, they do not utilize the information on Twitter. Furthermore, the content of most implied papers is not public when *ST tweets* are posted, so content is not a good factor to help predict scientific impact in real time.

2.2 Social Media Scientific Analysis

While most of the previous work focuses on structured data sources, there is some work focus on tweets. [28] explored the feasibility of measuring social impact and public attention to scholarly articles by analyzing buzz in social media. They explored the dynamics, content, and timing of tweets relative to the publication of a scholarly article, as well as whether these metrics are sensitive and specific enough to predict highly cited articles.

[29] studied approaches for defining and measuring information flows within tweets during scientific conferences. They suggest refinements of analyzing datasets based on tweets collected during scientific conferences and present our results from applying novel forms of intellectual tweet content analyses.

Many papers have only zero or one tweet mentioned, how to restrict the impact analysis on only those journals producing a considerable Twitter impact is a problem. [2] defined the Twitter Index (TI) containing journals with at least 80% of the papers with at least one tweet each. For all papers in each TI journal, they calculated normalized Twitter percentiles (TP) which range from 0 (no impact) to 100 (highest impact).

The approaches mentioned above are not appropriate for real-time prediction in Twitter because the formation of citation networks is time-consuming. In this paper, we use h-index instead of citation number as metric to evaluate the scientific impact and convert the traditional regression problem on citation number to a classification problem on h-index.

3 Overview

We look deeply into the tweets from our "paper accepted" data stream and find that some tweets are *NST tweets*. For example, the tweet: "*can the bank accept the toilet paper issued as by @UKenyatta as collateral??*", is a *NST tweet*, because the word "*paper*" means anything but a thesis in that tweet. Thus we need to build a classification model to filter *ST tweets*.

To predict scientific impact in Twitter in real time, we want to make use of the metadata of papers. It is surprising to find that *ST tweets* are always consisted of text blocks indicating metadata, such as authors and titles of papers and names, time and places of venues. We call these text blocks *Twitter Scholar Blocks* (*TSBs*) and build a sequence tagger to extract them.

We build a binary classification model to predict scientific impact in Twitter in real time. We use the model to judge whether the paper implied in *ST tweet* will increase the *h*-index [9] of its first author after five years. The *ST tweets* that imply accepted papers that will increase the *h*-indices of the first authors after five years are called High Impact Scholarly Tweets (*HIST tweets*). *TSBs* and information in Twitter social networks are combined in our model to predict *HIST tweets*. The framework of our approach is shown in Fig. 2.

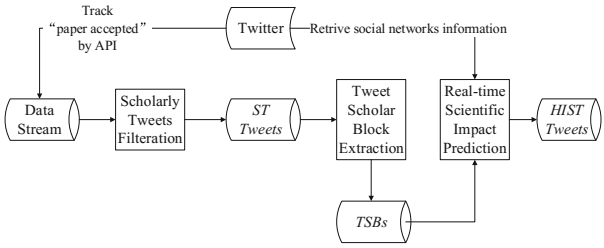

Fig. 2. Framework of our approach

4 Scholarly Tweets Filtration

We take filtering *ST tweets* from the data stream as a classification problem. To solve this problem, we propose an approach called Scholarly Tweets Filtration (*STF*) and build a classification model based on support vector machine (*SVM*) [3,4]. It is not easy to resolve the problem since there are two types of *ST tweets*. Some of them have specific paper titles, while others do not. For example, the tweet: "*Our paper 'Latent Space Model for Multi-Modal Social Data' accepted at @www2016ca #www2016!*", has an explicit title "*Latent Space Model for Multi-Modal Social Data*" surrounded by a pair of double quotation marks, while the tweet: "*New paper accepted!*", does not. The features we exploit are listed in Table 1.

To capture the information in social networks, the following features related to the author of the tweet are designed: **user's scholarly membership of academic institutions**. Obviously and empirically, the members of academic institutions have the higher probabilities to mention accepted papers. For simplicity, we collect names of academic institutions from the Internet and make a list containing the top sixty high-frequency words, such as "university", "institute", "research". Then we examine whether user descriptions contain words in the list to judge the existence of scholarly memberships.

Table 1. Features exploited in scholarly tweets filtration

Feature	Description
Scholar	Is the user a scholar
Bag-of-words	The bag of words of the tweet
Symbols	Words starting with symbols
Length	Text content length
Sentiment	The sentiment of tweet

To capture the information in tweet text, the following features are designed: **bag of words, words with trending symbols (e.g., "#", "*"), length of the tweet and the sentiment label of the tweet**. Words with trending symbols are commonly used to express topics on Twitter. In *ST tweets*, topics are often abbreviations of conferences, journals and research fields. We think the sentiment label is significant because our intuition is that no one would hide her happiness if her paper were accepted. Previous work shows sentiment analysis in citation context helps achieve better result [27]. The result of sentiment analysis is one of the three labels: positive, neutral and negative, according to our statistics, few of *ST tweets* is negative. In the experiment we used a free and open source tweet-specified sentiment analysis API[4] to generate sentiment labels for tweets.

To evaluate our *STF*, we manually labeled 5,400 tweets from our "paper accepted" data stream, nearly 45% are *ST tweets* and the ratio between *ST tweets* with and without explicit title is 10:7. Five-fold cross-validation was used in this experiment. FastText [10] was chosen as our baseline. The architecture of fastText is similar to the CBOW model [17], and it utilizes hierarchical softmax to reduce time expenditure. By training with SVM, the accuracies of *STF* and baseline are listed in Table 2. The performance of *STF* is 5.96% higher than the performance of the baseline. Although fastText model uses bag of n-gram as additional features to capture some partial information about the local word order, it only focuses on the text content. Thus the assistance of social features might improve the performance. The performance on *ST tweets* with explicit titles is 35.62% higher than the performance on *ST tweets* without explicit titles, which confirms the difficulties to filter *ST tweets* without explicit titles.

Table 2. Results of scholarly tweets filtration

Tweets	Number	***STF***	Baseline
All	5400	86.26%	81.37%
With titles	1429	98.04%	97.27%
No titles	1001	72.22%	70.43%

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

5 Tweet Scholar Block Extraction

To help predict scientific impact in Twitter in real time, we want to extract metadata from the implied papers. [13] found that a series of conventions allow users to tweet in structural ways using the combination of different blocks of texts which are combinations of plain text, hashtags, links, mentions and so on. We investigate that researchers post *ST tweets* also in structural ways using combinations of different Tweet Scholar Blocks (*TSBs*). Each *TSB* carries a part of meta data. Furthermore, the combinations of *TSBs* encode scholarly information about papers, venues, and authors. Six types of *TSB* are proposed by us. A *ST tweet* consisted of different types of *TSBs* is illustrated in Fig. 3 and every underlined sequence of tokens shows a type of *TSB*.

$$\left(\frac{\text{Dr Ramon Harvey}}{\text{Author}}\right)\ \left(\frac{\text{: My paper ”}}{\text{Other}}\right)$$

$$\left(\frac{\text{The Quert for Qist : Defining Societal Justice in the Qur'an}}{\text{Title}}\right)$$

$$\left(\frac{\text{” has been accepted for the}}{\text{Other}}\right)\ \left(\frac{\text{BRAIS conference}}{\text{Venue}}\right)\ \left(\frac{\text{on}}{\text{Other}}\right)$$

$$\left(\frac{\text{11–12 April 16}}{\text{Time}}\right)\ \left(\frac{\text{London}}{\text{Place}}\right)\ \left(\frac{\text{.}}{\text{Other}}\right)$$

Fig. 3. An example of tweet scholar blocks

Author: The names of authors (e.g., Dr. Ramon Harvey).
Title: The title of the paper (e.g., The Quert for Qist: Defining Societal Justice in the Qur'an).
Venue: The short name of the venue (e.g., BRAIS conference) or its entire name (e.g., The British Association for Islamic Studies conference).
Time: The time expression when the venue will be held (e.g., 11–12 April 16).
Place: The place where the venue will be held (e.g., London).
Other: The rest part of tweet text that does not belong to the above five types.

In order to test the validity that *ST tweets* are constructed by different combinations of *TSBs*, we randomly chose 1,400 *ST tweets*. Firstly, we used an annotator[5] [8,20] to tokenize those tweets. Secondly, we manually labeled *TSBs* in BIO schema [23]. The *ST tweets* that are consisted of only *Other* type of *TSBs* are 17.73%. It means most *ST tweets* contain at least one non-*Other* type of *TSBs*. Therefore, we can extract some metadata from papers by extracting *TSBs*.

[5] http://www.cs.cmu.edu/~ark/TweetNLP/.

We regard extracting *TSBs* from *ST tweets* as a sequence labeling problem. To solve this problem, we propose an approach to extract *TSBs* called Tweet Scholar Block Extraction (*TSBE*) and build a sequence tagger based on conditional random fields (*CRFs*) [11]. Due to the informal and short nature of the tweets, we apply a tweet-specified POS tagger which is the same annotator we use to tokenize to produce POS labels. We use the tweet NER tagger [24,25] to extract features for *Time* and *Place* types of *TSBs*.

By analyzing *ST tweets*, we discover that most of the Twitter account mentioned in *ST tweets* are co-authors. So we use tokens starting with "@" (e.g., @LiuQunMTtoDeath) in *ST tweets* to help extract *Author* type of *TSBs*.

It is also investigated that the titles of papers usually occupy up to 40% of the text content which are often surrounded by pairwise symbols (e.g., "" and ‘’) or all capitalized to show their differences. So to extract *Title* type of *TSBs*, we judge whether a token is capitalized or surrounded by pairwise symbols.

In *ST tweets*, nearly 87% words with trending symbols indicate venues (e.g., #SPAFACON2016). Some Twitter accounts mentioned (e.g., @acl2016) and preposition phrases (e.g., *by the Intl Jrl of Osteopathic Medicine*, *in Chem. Sci.*) also mean venues. Therefore, we use words with trending symbols and prepositions to help extract *Venue* type of *TSBs*.

Table 3. Statistics of the five none-*other* types of *TSBs*

Type	Number	Precision	Recall	F1 Value
Title	304	93.55%	74.36%	82.86%
Author	732	82.81%	72.60%	77.37%
Place	217	79.23%	72.94%	75.96%
Venue	973	83.12%	72.73%	77.58%
Time	235	77.40%	74.55%	75.95%

To evaluate our *TSBE*, we used the 1,400 *ST tweets* described above to train and test. Five-fold cross-validation was used in this experiment. The precision rates, recall rates and F1 values of the five none-*Other* types of *TSBs* are shown in Table 3. The *Other* type of *TSBs* is neglected because the label of this kind is "O" in BIO schema and we do not care in this paper. The performance of *Title* type of *TSBs* is the highest. Judging whether a token is capitalized or surrounded by pairwise symbols might be useful features. The performances of *Place* and *Time* types of *TSBs* are close. And the performances of *Author* and *Venue* types of *TSBs* are also close. It might be that they are strongly correlated to the representation of the leading symbols and prepositions, so similar features take effects. We analyze some blocks with wrong predictions and find that some *Time* and *Place* types of *TSBs* were affected by the errors produced in the tweet NER tagger. To improve our performances, we need to decrease the errors produced in the tweet NER tagger and enhance the representation of leading symbols and prepositions.

6 Real-Time Scientific Impact Prediction

We regard predicting scientific impact in Twitter in real time as a classification problem of judging whether the paper implied in *ST tweet* will increase the h-index of its first author after five years. To solve this problem, we propose an approach called Real-time Scientific Impact Prediction (*RSIP*) and build a classification model based on support vector machine (*SVM*). Previous work shows that the scientific citation process acts relatively independently of the social dynamics on Twitter [30], so we take both social networks information in Twitter and text information generated from *TSBs* into account. As the paper implied in *ST tweets* may be not public, we can not use its content. Thus we can only think of metadata of articles. We categorize the features we exploit into three categories: author social features, venue popularity features, and title features. The features we exploit is listed in Table 4.

Table 4. Features exploited in real-time scientific impact prediction

Feature	Description
Sum Friends Count	Sum friends number of all the authors
Sum Followers Count	Sum followers number of all the authors
Sum Statuses Count	Sum statuses number of all the authors
Maximum Friends Count	Maximum friends number of all the authors
Maximum Followers Count	Maximum followers number of all the authors
Maximum Statuses Count	Maximum statuses number of all the authors
Minimum Friends Count	Minimum friends number of all the authors
Minimum Followers Count	Minimum followers number of all the authors
Minimum Statuses Count	Minimum statuses number of all the authors
Average Friends Count	Average friends number of all the authors
Average Followers Count	Average followers number of all the authors
Average Statuses Count	Average statuses number of all the authors
Friends Count	Friends number of the first author
Followers Count	Followers number of the first author
Statuses Count	Statuses number of the first author
Individual Verification Status	Is the first author verified
Group Verification Status	Is anyone among all the authors verified
Retweets Count	Retweets number of the *ST tweet*
Replies Count	Replies number of the *ST tweet*
Liked Count	Liked number of the *ST tweet*
Current Tweets Count	Tweets number in the venue
History Tweets Count	Average history tweets number in the venue
Title Embedding	Sum of word embeddings in title

6.1 Author Social Features

To capture the author social information, we try to find reasonable representations of influences of authors. It is investigated that the first author usually leads the collaboration. Besides, previous work shows that the overall impact of all co-authors should have the potential to influence a paper's quality and popularity, which will further affect a researcher's h-index [6]. We use the *Author* type of *TSBs* extracted to find the authors in *ST tweets*. And the first author is defined as follows. If the *ST tweet* is original, its author is the first author of the paper. Otherwise, the first *Author* type of *TSB* indicates the first author of the paper. We think the influence of an individual is related to her friends number, followers number, statuses number. To show the influence of a group, we calculate the sum, maximum value, minimum value and average value of influences of the participants in that group. In spite of these, we take statuses of user verification into account. Verification is used by Twitter mostly to confirm the authenticity of celebrity accounts. Previous work found that 91% of tweets written by verified users are retweeted, compared with 6% of tweets where the author is not verified [21], which means that tweets posted by celebrities are more popular. Additionally, we calculate retweets count, replies count and liked count of the *ST tweet*.

6.2 Venue Popularity Features

Google Scholar metrics[6] shows that different venues have large differences in their *h5*-indices (the h-index when only considering articles published within the last 5 complete years). Since the well-respected venues are better platforms for researchers to publish their work or results, our intuition is that top sites help scholars spread their scientific impact. And increasing the citation counts of their papers further offers an enormous potential to increase their h-indices.

Scholars often use Twitter as a note-taking tool [14] during venues, so the tweets number in the venue topic may reflect the popularity and influence of the site. We use the quantity of statuses in the venue topic to represent the popularity of the venue. Considering the developments and the trends of the venues, we also take the historical total quantity of statuses into account.

6.3 Title Features

Every scientific paper has its specified topics, while the popularity of topics may influence the speed of the appearance of scientific impact [1]. We think the title is the most direct and attractive way to declare research topic.

To capture the influence of topics, we attempt to extract information from learning representations of the titles of scientific papers. To learn a good representation of the titles of scientific papers, we first use word2vec

[6] https://scholar.google.com/citations?view_op=top_venues.

[17–19] and pre-trained 300-dimensional word embedding *GoogleNews-vectors-negative300.bin.gz*[7] which is trained from the Google News corpus to obtain the representations of words, and then we sum up all the corresponding embeddings of the words in title split with whitespaces. If there is no explicit title in the *ST tweets*, we set the Title Embedding all zeros to make this feature not work.

7 Experiments

To evaluate our approaches, firstly, we manually labeled tweets from our "paper accepted" data stream and set experiments to compare the performances between *RSIP* and the baseline. Then we did feature selection experiment to find the best feature conjunction of *RSIP* to improve performance. At last, we did feature analysis experiment to find the effectiveness of each feature in the best feature conjunction of *RSIP* and which features, in particular, are highly valued. Five-fold cross-validation was used in our experiments. Accuracy was used as the evaluation metric.

7.1 Data Set

We randomly chose 273 *ST tweets* posted in 2011 from our "paper accepted" data stream and found the true names of authors, titles of papers, names, time and places of venues by a scholarly search engine such as Google Scholar[8]. There are no *NST tweets* in the data set, since *NST tweets* do not imply accepted papers and it is meaningless to feed them into the baseline we chose. In these *ST tweets*, 142 *ST tweets* of them are without specific titles, while others are with explicit titles. According to the information we found, we then use Google Scholar to gather h-index of every first author and citation number of every corresponding paper in 2016. Now we can know whether the papers accepted in 2011 will increase h-indices of their first authors after five years in 2016. If the paper's citation number in 2016 is more substantial than its author's h-index in 2016, it means the article accepted in 2011 increased its primary author's h-index after five years in 2016. In such way, we manually labeled the data.

7.2 Real-Time Scientific Impact Prediction Evaluation

Since there are few related works about scientific impact prediction in real time, we took the approach of [5], which is the state-of-the-art method to predict the scientific impact on citation networks, as our baseline to simulate real-time prediction. In the baseline, temporal and topological features derived from citation networks are used to predict a paper's future impact (e.g., number of citations). The baseline uses a behavioral modeling approach in which the temporal change

[7] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit?usp=sharing.

[8] https://scholar.google.com.

in the number of citations a paper gets is clustered, and new papers are evaluated accordingly. Then, within each cluster, the impact prediction is modeled as a regression problem where the objective is to predict the number of citations that a paper will get in the near or far future, given the early citation performance of the paper. The baseline produced the citation number of each paper in our dataset after five years. And we compared the citation numbers to the real first authors' h-indices in 2016 to judge whether the papers increased its first author's h-index in 2016.

Table 5. Comparing results between baseline and *RSIP*

Approach	Accuracy
Baseline	63.00%
TSBE+RSIP	73.99%
RSIP	78.02%

We compared the result of using *TSBE* and *RSIP* (*TSBE+RSIP*) with the result of using *RSIP* on manually labeled *TSBs* and the result of the baseline. Results are shown in Table 5. Overall, it is feasible to predict scientific impact in Twitter in real time. The performance of *RSIP* is higher than the performance of the baseline. The reason might be that the baseline is not appropriate for predicting scientific impact in real time. And the performance of *TSBE+RSIP* is lower than the performance of *RSIP* on manually labeled TSBs. The errors produced in *TSBE* might affect the performance of *RSIP*.

7.3 Feature Selection

To find the best feature conjunction of the features to improve the performance of our real-time scientific impact prediction model, we used an advanced greedy feature selection method the same as [7] used. Figure 4 shows the feature selection approach mentioned above.

Since greedy feature selection approach suffers from data sparseness, it is always blocked by a local optimum feature set. To find a global optimum feature set, this approach uses random techniques to generate several feature sets first and then run greedy feature selection on the best one among them. Finally, we find that the best feature conjunction consisted of *Sum Friends Count*, *Sum Followers Count*, *Followers Count* and *Title Embedding*. We call it *RSIP_Best*.

Results in Table 6 illustrate that the best feature conjunction (*RSIP_Best*) outperforms *RSIP* by about 3.76% on our manually labeled dataset. The three kinds of features, namely maximum, minimum, average counts are not in the best feature conjunction. It might be that these three kinds of counts do not reflect the entire influences of groups and are often limited by the variances of individual authorities. Additionally, the performance of *TSBE+RSIP_Best* is lower than the performance of *RSIP_Best* and is 4.96% higher than the performance of *TSBE+RSIP*. It might be that the errors produced in *TSBE* effect *RSIP_Best*.

An advanced greedy feature selection algorithm.
Input: All features we extracted.
Output: the best feature conjunction *BFC*
Procedure:
Step1: Randomly generate 80 feature set *F*.
Step 2: Evaluate every feature set in *F* and select the best one denoted by *RBF*.
Features excluded those in *RBF* are denoted as *EX_RBF*
Step 3: t = 0,*BFC*(t)=*RBF*;
  Repeat
    Foreach feature in *EX_RBF*
      If Evaluation(*BFC*)
        < Evaluation(*BFC*, feature)
        BFC(t+1) = {*BFC*(t), feature}
        EX_RBF(t+1) = *EX_RBF*(t) – {feature}
  While *BFC*(t+1) ≠ *BFC*(t)
Note: Evaluation(*BFC*) refers to the performance of ranking function trained from features in *BFC* on validation data.

Fig. 4. Advanced greedy feature selection algorithm used in feature selection

Table 6. Comparing results with best feature conjunction

Approach	Accuracy
RSIP	78.02%
RSIP_Best	80.95%
TSBE+RSIP	73.99%
TSBE+RSIP_Best	77.66%

7.4 Ablation Study

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

By comparing the results shown in Table 7, we can see that *Sum Followers Count* is very effective to our *RSIP_Best*. The reason might be that *Sum Followers Count* is more suitable to stand for the influence of the authors' group.

Meanwhile, *Title Embedding* is not so efficient in our data. Perhaps the reason is that 52.01% of the *ST tweets* do not have specific titles. So the feature only works on the rest *ST tweets*.

Table 7. Comparing results by decaying every feature one by one

Approach	Accuracy
RSIP_Best	80.95%
RSIP_Best-Sum Friends Count	75.09%
RSIP_Best-Sum Followers Count	73.99%
RSIP_Best-Followers Count	76.56%
RSIP_Best-Title Embedding	79.85%
TSBE+RSIP_Best	77.66%
TSBE+RSIP_Best-Sum Friends Count	69.60%
TSBE+RSIP_Best-Sum Followers Count	65.93%
TSBE+RSIP_Best-Followers Count	68.86%
TSBE+RSIP_Best-Title Embedding	73.63%

8 Conclusion

In this paper, we propose *STF*, *TSBE* and *RSIP* to predict scientific impact in real time. The accuracy of *RSIP_Best* is 80.95%, which outperforms the baseline based on citation networks.

The best feature conjunction consists of the sum friends and followers count of all the co-authors, followers count of the first author and title embeddings. And sum followers count of all the co-authors is the most critical feature. The results show that Twitter has the potential to predict scientific impact in real time and our novel approach can achieve comparable performance. Hope real-time scientific impact prediction in Twitter can help researchers to expand their influences and more conveniently "stand on the shoulders of giants".

Acknowledgments. We very appreciate the comments from anonymous reviewers which will help further improve our work. This work is supported by National Natural Science Foundation of China (No. 61602490).

References

1. Bethard, S., Jurafsky, D.: Who should I cite: learning literature search models from citation behavior. In: CIKM, pp. 609–618 (2010)
2. Bornmann, L., Haunschild, R.: How to normalize Twitter counts? A first attempt based on journals in the Twitter index. Scientometrics **107**, 1405–1422 (2016)
3. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers, pp. 144–152 (1992)
4. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
5. Davletov, F., Aydin, A.S., Cakmak, A.: High impact academic paper prediction using temporal and topological features. In: CIKM, pp. 491–498 (2014)

6. Dong, Y., Johnson, R.A., Chawla, N.V.: Will this paper increase your h-index?: Scientific impact prediction. In: WSDM, pp. 149–158 (2015)
7. Duan, Y., Jiang, L., Qin, T., Zhou, M., Shum, H.: An empirical study on learning to rank of tweets. In: COLING, pp. 295–303 (2010)
8. Gimpel, K., et al.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers, pp. 42–47 (2011)
9. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A. **102**(46), 16569–16572 (2005)
10. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: EACL, pp. 427–431 (2017)
11. Lafferty, J., Mccallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data, pp. 282–289 (2001)
12. Letierce, J., Passant, A., Breslin, J.G., Decker, S.: Using Twitter during an academic conference: the iswc2009 use-case. In: ICWSM, pp. 279–282 (2010)
13. Luo, Z., Osborne, M., Petrovic, S., Wang, T.: Improving Twitter retrieval by exploiting structural information. In: AAAI, pp. 648–654 (2012)
14. Mapes, K.: A qualitative content analysis of 19,000 medieval studies conference tweets. In: ACM International Conference on the Design of Communication, p. 48 (2016)
15. Mckeown, K., et al.: Predicting the impact of scientific concepts using full text features. J. Assoc. Inf. Sci. Technol. **67**, 2684–2696 (2015)
16. McNamara, D., Wong, P., Christen, P., Ng, K.S.: Predicting high impact academic papers using citation network features. In: Li, J., et al. (eds.) PAKDD 2013. LNCS (LNAI), vol. 7867, pp. 14–25. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40319-4_2
17. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS, pp. 3111–3119 (2013)
19. Mikolov, T., tau Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: HLT-NAACL, pp. 746–751 (2013)
20. Owoputi, O., O'Connor, B.T., Dyer, C., Gimpel, K., Schneider, N., Smith, N.A.: Improved part-of-speech tagging for online conversational text with word clusters. In: HLT-NAACL, pp. 380–390 (2013)
21. Petrovic, S., Osborne, M., Lavrenko, V.: RT to win! predicting message propagation in Twitter. In: ICWSM, pp. 586–589 (2011)
22. Priem, J., Costello, K.L.: How and why scholars cite on Twitter. Proc. Asist Ann. Meet. **47**(1), 1–4 (2010)
23. Ratinov, L.A., Roth, D.: Design challenges and misconceptions in named entity recognition. In: CoNLL, pp. 147–155 (2009)
24. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: an experimental study. In: EMNLP, pp. 1524–1534 (2011)
25. Ritter, A., Mausam, Etzioni, O., Clark, S.: Open domain event extraction from Twitter. In: KDD, pp. 1104–1112 (2012)
26. Shibata, N., Kajikawa, Y., Matsushima, K.: Topological analysis of citation networks to discover the future core articles. JASIST **58**, 872–882 (2007)
27. Small, H.G.: Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics **87**, 373–388 (2011)

28. Thelwall, M., Priem, J., Eysenbach, G.: Can tweets predict citations? Metrics of social impact based on twitter and correlation with traditional metrics of scientific impact. J. Med. Internet Res. **13**, e123 (2011)
29. Weller, K., Dröge, E., Puschmann, C.: Citation analysis in Twitter: approaches for defining and measuring information flows within tweets during scientific conferences. In: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages, Heraklion, Crete, Greece, 30 May 2011, pp. 1–12 (2011)
30. de Winter, J.C.F.: The relationship between tweets, citations, and article views for plos one articles. Scientometrics **102**, 1773–1779 (2014)
31. Yogatama, D., Heilman, M., 'connor, B.O., Dyer, C., Routledge, B.R., Smith, N.A.: Predicting responses and discovering social factors in scientific literature predicting responses and discovering social factors in scientific literature (2011)
32. Yogatama, D., Heilman, M., O'Connor, B.T., Dyer, C., Routledge, B.R., Smith, N.A.: Predicting a scientific community's response to an article. In: EMNLP, pp. 594–604 (2011)