Predicting User Influence in Social Media

Chunjing Xiao 1,*, Yuhong Zhang 2, Xue Zeng 1, and Yue Wu 1
1. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
2. College of Information Science and Engineering, Henan University of Technology, Zhengzhou, China
Email: chunjingxiao@gmail.com, yhzhang@haut.edu.cn, xuezeng68@gmail.com, ywu@uestc.edu.cn
*Corresponding author

Abstract—Understanding influence plays a vital role in enhancing businesses operation and improving effect of information propagation. Therefore the user influence in social media, such as Twitter, is widely studied based on different standards, such as the number of followers, retweets and so on. However, little work considers the accurate click number of short URLs as the measurement of influence. In Twitter short URLs are frequently included in tweets because of the limitation of characters. And some users may focus more on click number of the URLs instead of the number of followers or retweets. Thus, it is necessary to analyze the factors that impact the click number received by URLs of users. In this paper, we conduct the predictive analyses about the user influence which is measured by the click number of short URLs. We first exploit a wide range of possible features consisting of the sets of user properties, behavior and topics. And then we employ the logistic regression analysis to identify the significant features for predicting the user influence, and find most of features we proposed have a significant predictive power to the user influence. Finally based on the large scale Twitter data, four models are used for the prediction and the Bagging model achieves the best result, an overall accuracy of more than 82%.

Index Terms—Twitter; Influence; Web Traffic; Predict

1. INTRODUCTION

Social Media such as Twitter and Facebook has become an important platform to publish or receive information, which is changing the way of communication and knowledge sharing between the people. Users in these systems post and discussion millions of news, options, and reviews to promote them. Correspondingly, influence, which has long been studied in the fields of sociology, communication, marketing and political science [1, 2], also receive much attention in social media, because understanding influence can provide insights for users to learn why certain information propagates faster and how we improve the effect of contents diffusion.

Currently, the user influence has been analyzed from different aspects based on different standards. For example, Cha et al. [3] present a comparison of three measures of influence: the number of user followers, retweets and mentions in Twitter. Also, Kwak et al. [4] analyze the user influence based on another three standards: the number of followers, PageRank and retweets. Subsequently, the amount of extended retweets is used as the standard for predicting the influence of Twitter users [5], and for analyzing word-of-mouth information propagation in Twitter [6]. Whereas none of these works consider the accurate click number of short URLs as the standard of influence. In fact shortened URLs are frequently included in the content published by users, because of the limitation of characters of contents, especially for Twitter which limits a tweet to 140 characters. And there should be a lot of users who aim to attract web traffic by Twitter. In addition, in Twitter the number of retweets and click number of URLs received by the user are disproportionate [7]. Therefore it is necessary to understand user influence based on the standard of click number of short URLs.

In this paper, we predict the user influence based on the accurate click number of short URLs. Due to the importance of click number, Antoniades et al. [8] compare the popular websites in Twitter, which are measured by the click number received via Twitter, with that in Alexa.com. And Romero et al. [9] use the global click number of URLs as the ground truth to evaluate their proposed algorithm of ranking users in Twitter, but as they said the global clicks are noise because they also include the click source from outside of Twitter, such as Facebook and forums.

While compared with the existed work, we use the accurate click number as the standard of user influence. According to the click information provided by Bitly.com, there are three kinds of click number: accurate clicks referring to the click number received by each URL of each user; domain clicks referring to the click number received by any object in a domain (such as Twitter.com); global clicks referring to the sum of all the click number. Therefore the accurate click number is a precise one without noise in comparison with the domain clicks and global clicks.

Based on the standard of the accurate click number, we predict the influence of users in Twitter. And we treat the prediction as a classification task by defining four categories to represent the levels of the user influence. To conduct the prediction, we exploit a wide range of possible features consisting of the sets of user properties, behavior and topics. The set of user properties includes the basic properties of users, such as the number of followers, friends and lists, as well as the properties we...
defined, such as the number of active followers and the type of user domains. The set of behavior is composed of the number of tweets and the entropies of published time of URLs. And the set of topics includes the topic category and topic entropy. After extracting these features, we first analyze the significance of each feature by using the logistic regression analysis and find majority of features are statistically significant. And then by using multiple classification models, we predict the levels of the user influence and achieve an overall accuracy of more than 82% with the Bagging model.

II. RELATED WORKS

The studies related to influence in online social networks have been conducted from ranking influential users [3, 4, 10, 11], and quantifying user influence [5, 9] to predicting popularity of contents [12-15]. Specifically, Kwak et al. [4] find that the ranking of users depending on the amount of retweets is different from that depending on the number of followers and PageRank in the follower network. And Cha et al. [3] also compare the user influence based on indegree (the number of followers), the number of retweets and number of mentions, and demonstrated that popular users with high indegree are not necessarily influential in terms of spawning retweets or mentions. Besides, influential users are identified in Twitter by taking the topical similarity and the link structure into account [10] and by using modified k-shell decomposition algorithm [11].

Apart from ranking influential users, Hofman et al. [5] quantify influence of general users based on the standard of the number of extended retweets. This standard, beyond official retweets, also includes the amount that implicit propagation which will occurs when a user shares a URL that has already been shared by one of his friends (followings) without necessarily citing the information source. Based on this measure, to predict user influence they explore features consisting of the numbers of followers, friends and tweets, date of joining, and past influence of users including past total influence and local influence. Their predictive model, the regression tree, achieves relatively poor performance ($R^2 = 0.34$) without averaging predicted and actual values at the leaf nodes. And since the majority of users act as passive information consumers and do not forward the content to the network, Romero et al. [9] developed an efficient algorithm to quantify the influence of all the users in Twitter by taking passivity into account, and they used the global clicks of shorts URLs as the standard of the influence, which is noise as they said.

Another body of works is the prediction of the popularity in Twitter. Hong et al. [12] predict the popularity of tweets as measured by the number of future retweets. They define several categories to represent the volume of retweets and predict which categories the tweets will belong to. The prediction that whether a tweet will be retweeted is studied in [13]. Based on the model with the passive-aggressive algorithm, they can automatically predict retweets and find that the performance is dominated by social features, but the tweet features add a substantial boost. Artzi et al. [14] predict whether a message will elicit a user response in Twitter based on a discriminative model, and they explore various sources as features, such as the language used in the tweet, the user's social network and history. Bandari et al. [15] predict the popularity of news items in Twitter prior to their release. A multi-dimensional feature space derived from properties of the article is exploited for the prediction.

However, differing from these studies, we use the different standard for the user influence, the accurate click number of short URLs. Compared with the existed work, the analyses and prediction based on click number can provide insights to improve the web traffic via social media. In addition, we explore different features for the prediction, such as the type of tweets, the entropy of published time of URLs and so on.

III. DATA DESCRIPTION

As our goal is to predict user influence measured by the accurate click number of short URLs, the data for the experiments should be mainly comprised of the information of short URLs in tweets published by Twitter users and accurate click number received by these short URLs.

To obtain these data, we firstly select the targeted users of Twitter. Particularly, we select users who tend to publish tweets including short URLs, and these URLs should be hosted by Bitly because the short URLs of Bitly take up about 50% of all the URLs in Twitter [8] and their accurate click number can be collected.

To this end, from more than 790 millions tweets during June 2012 collected by Twitter streaming APIs which return roughly 10% of all public tweets, we extract around 46 million unique users. From these users, we select users satisfying the following conditions: (i) The language in the profile settings of users is English, because users speaking English are most popular in Twitter [16] and we are familiar with this language; (ii) The ratios between the numbers of tweets including Bitly URLs and all the tweets of users are no less than 80%, because, for users with many Bitly URLs, their influence can more properly be represented by accurate click number of their URLs; (iii) The domain focuses of users are more than 80%, and focused domains are the same with domains of websites in their profiles. Here the domain focus refers to the highest fraction of the number of URLs of a domain over the entire number of URLs, and is defined as below:

$$D_i = \frac{1}{V_i} \max_{k} v_{ik}.$$  \hspace{1cm} (1)$$

where $V_i$ refers to the sum of URL number of user $i$, and $v_{ik}$ refers to the number of URLs with domain $k$ of user $i$. We employ this selection because this kind of users are more likely to aim to attract the web traffic via Twitter; (iv) Users publish at least average one URL per day, because it is obvious that too few URLs will skew the results. As a result, 32,942 users are selected as our targeted users.
And then, for these selected users, by Twitter APIs we download their profiles, followers and the lists that include them. And there are more than 194.69 million follower links and 4.33 million list links. We also download their tweets during June by Twitter APIs, around 9.13 million. Among them, approximately 8.57 million tweets include the short URL, and the click information of these URLs is downloaded by Bitly APIs. The detailed information is presented in Table I.

<table>
<thead>
<tr>
<th>TABLE I. TWITTER DATA DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of users</td>
</tr>
<tr>
<td>Number of tweets</td>
</tr>
<tr>
<td>Number of tweets with Bitly URLs</td>
</tr>
<tr>
<td>Number of follower links</td>
</tr>
<tr>
<td>Number of lists links</td>
</tr>
</tbody>
</table>

IV. FEATURES ENGINEERING

Here we introduce the features that will be used to predict user influence. We try to explore a wide range of possible features which help determine the attributes related to user influence. The features consist of the sets of user properties, user behavior and topics.

A. Features of User Properties

We first describe the features about user properties. Based on the user information we can collect, the metadata, such as the number of followers, friends and lists, will be extracted as the feature. Besides we exploit relative information to further describe user characters, such as the number of active followers and user domains.

1) Number of followers: Followers of a user are the people who will receive all the updates or messages published by this user. And the number of followers of a user can directly indicate the size of the audience for this user. Therefore, the number of followers is frequently used as measuring the user influence [3, 4]. Hence, this number will be extracted as the feature to predict user influence in this paper. Correspondingly, another basic property of users, the number of friends which can reflect the social capital, will also be as a feature.

In addition, because there are a lot of accounts that have been suspended due to spammers or other reasons [17], and there should be a part of users that register multiple accounts and only use one of them or stop using Twitter. We further compute the number of active followers as the feature for prediction. To this end, we need identify whether a user is active. In general, the users will be regarded as active ones by the owners of online social networks, such as Twitter or Facebook, if they log in at least once a month [18]. However, since we cannot obtain the information about logging in, we regard users as active ones if they publish at least one tweet in recent two months. After collecting the recent tweets of each follower by Twitter APIs, we can compute the number of active followers for each user.

2) Number of Lists: The Twitter List, launched on November 2009, is an official functionality to group sets of users into topical or other categories, and it aid to help users organize people they follow. If a user has been added into more lists, it means that this user receive more attention from others. Therefore, the number of lists including a user, to some extent, should reflect the popularity of this user.

We also analyze the correlation between the number of lists and the click number of URLs of users in Fig. 1. The x-axis is the number of lists into which the users have been added, and y-axis refers to the average click number of URLs published by the users. From this figure, we can see that the correlation exhibits some linear characteristic, and the linear correlation coefficient is 0.7337. This indicates that the number of lists cannot accurately reflect the click number, however there exist a certain linear relationship between them. Hence, we explore the number of lists as a feature to identify its importance in the predictive model.

3) User domains: The short URLs can be generated by users’ own domains or public domains provided by companies of short URLs, such as Bitly or Ownly. We try to find whether the special domains of short URLs have a significant impact on the click number by exploring the feature of user domains. To compute this feature, we need learn whether a domain is the special one or public one. To this end, we identify public domains by check whether their short URLs are extended to long URLs with multiple domains, i.e., the domains of short URLs will be regarded as the public ones if the corresponding long URLs are directed to multiple domains. As a result, two domains, bit.ly and j.mp, are identified as public ones and others as special ones. Consequently based on the predominant domains of short URLs published by users we can classify users into two groups: one with special domains and another one with public domains.

B. Features of User Behavior

Compared with the features of user properties, the features of user behavior mainly describe characters which can be controlled arbitrarily by users. For example, the type of tweets and the published time of tweets can be easily changed by users. Therefore analyzing which features have a predictive power for user influence is important for users to adopt their behavior.

1) Type of tweets: The Twitter provides different marks to enhance the contents of tweets, such as hashtags and mentions. The hashtag, whose format is #keyword, can be a kind of mark about keywords or topics of the tweet for convenience of searching or categorizing messages. And the mention, whose format is @username, will be a kind of the conversation between users of Twitter. The tweets

© 2013 ACADEMY PUBLISHER
might include only hashtags, only mentions, or both them. Correspondingly, the tweets can be called as hashtag tweets, mention tweets, or hashtag and mention tweets. To conduct prediction, for each kind of tweets, the ratio between the numbers of them and all the tweets will be calculated as the feature. Also the number of all the tweets is computed as a feature.

\[ \text{E}_R = \max \sum_{i=1}^M d_i \cdot \ln \frac{d_i}{M} \]  

(2)

where \( d_i \) refers to the number of URLs published during the \( h \) hour by the user \( i \), \( d_i \) is the sum of the number of URLs published by the user \( i \), and \( M \) is the sum of all the hours. For a user, if her tweets are published only in one hour, the time entropy will be 0; while if her tweets are published in the \( M \) hours uniformly, the time entropy will be 1. Hence, higher entropy denotes that users have the lower inter-tweet delays and tend to publish tweets regularly. We will compute the entropies for day time and night time as the two features.

2) Published time: To exploit the character of published time of tweets, we firstly analyze the number of URLs published in different hours, as shown in Fig. 2. The y-axis is the average number of URLs. The figure clearly shows that more tweets are published in the day and less in the night. Because of the significant difference in the day and night, for each user we divide his tweets into two groups: tweets in day time and tweets in night time. For convenience, we simply regard the day time as from 8 AM to 7 PM and others as the night time.

By intuition, there are two properties relative to time of URLs that might impact the click number: the amount of URLs and intervals of published time of URLs. For example, if a user publishes too many tweets in a short time, the large amount of information will be beyond the receptive ability of his audience, and a part of URLs will be skipped. Hence, we introduce a comprehensive variable, time entropy, to measure the number of URLs and intervals of published time of URLs as the feature of our predictive model. For user \( i \), his time entropy \( E_i \) is defined as below:

\[ E_i = -\frac{1}{\ln M} \sum_{h=1}^M \frac{d_i}{d_i} \cdot \ln \frac{d_i}{d_i} \]  

(3)

where \( d_i \) refers to the number of URLs published during the \( h \) hour by the user \( i \), \( d_i \) is the sum of the number of URLs published by the user \( i \), and \( M \) is the sum of all the hours. For a user, if her tweets are published only in one hour, the time entropy will be 0; while if her tweets are published in the \( M \) hours uniformly, the time entropy will be 1. Hence, higher entropy denotes that users have the lower inter-tweet delays and tend to publish tweets regularly. We will compute the entropies for day time and night time as the two features.

C. Features of Topics

Here we exploit the features relative to the topics. We want to measure whether the topic category and topic distribution in the tweets are important for impacting the user influence. Thus these features include two values: user topic category and topic entropy.

1) User topic category: The role of content is generally analyzed in the work related to influence in online social networks. For example, Cha et al. [3] study the propagation of three popular topics in 2009 in Twitter, and find that most influential users can hold significant influence over a variety of topics. And Hofman et al. [5] use humans to classify the content of a sample of 1000 URLs and find that the content features are not informative in predicting influence in Twitter. However, we use the different standard, accurate click number, to represent user influence. Besides, we classify users automatically into different topic categories. Therefore, here we exploit the feature of user topics for the prediction to measure the correlation between the accurate click number and the content of tweets.

To this end, we firstly need classify users into different topic categories. In Twitter, organizations and individuals tend to create multiple accounts for publishing different contents. For example, there are more than 30 accounts for Washington Post [19]. Thus, the accounts in Twitter can be divided into different topic categories. The method of the classification is mainly based on the Twitter list. Because the names and descriptions of Twitter lists provide valuable semantic cues to the experts’ domain of expertise [20], the names of lists can be used to classify users. Specifically, if a user is frequently added into lists with similar names, it will be put in the category related to these names. For example, if the New York Times is often added into lists with the name of News, it will be put in the News category. To compute the most frequent names of lists including a user, we define dominant frequency of list names for user \( i \), \( R_i \), as below:

\[ R_i = \max_{m} L_{im} \]  

(3)

where \( m \) is the number of names of lists that include user \( i \), and \( L_{im} \) is the number of lists with the \( m \)-th name. Correspondingly, the dominant name refers the name of the lists with dominant frequency.

Following the steps of the classification procedure, we first clean the data by removing users who are included in less than 10 lists and explore the stems of frequent names of lists. Second we select nine categories based on the data we downloaded: Tech, News, Music, Sports, Food, Politics, Education, Health and Travel, and compare the dominant names of lists which include the users with these nine categories. The users whose dominant names of lists can match one of the nine categories will be put into the corresponding category. Otherwise they will be excluded from any category.

2) Topic distribution: The topics may have a wide range in the tweets for some users while narrow for others. Even if multiple users belong to one topic category, their topic ranges might still be different. For example, for the two users from the news category, one has a wider range of topics if he publishes tweets including international news, domestic news and technology news, while another has a comparatively narrower one if he only publishes domestic news. To
measure the wide degree of topics, we use Latent Dirichlet Allocation (LDA) [21] to compute the topic distribution of users. LDA is an unsupervised machine learning generative probabilistic model, which identifies latent topic information in large collections of data including text corpora, and has been widely used to exploring the interest and topic [22, 23]. For a corpus of \( M \) documents, LDA assumes that documents are generated from a set of \( N \) latent topics. In a document, each word \( w_i \) is associated with a hidden variable \( z_i \in \{1, \ldots, N\} \) indicating the topic from which \( w_i \) is generated. The probability of word \( w_i \) is expressed as:

\[
P(w_i) = \sum_{j=1}^{N} P(w_i | z_j)P(z_j).
\]

where \( \beta = P(w_i | z_j) \) is a probability of word \( w_i \) in topic \( j \) and \( \theta = P(z_j = j) \) is a document specific mixture weight indicating the proportion of topic \( j \). LDA treats the multinomial parameters \( \beta \) and \( \theta \) as latent random variables sampled from a Dirichlet prior with hyper-parameters \( \alpha \) and \( \eta \) respectively.

To compute topic distribution, for each user, we first merge all its tweets into one document. And similar to [22], we remove the 570 stop-words and terms that cannot be found in Wikipedia dataset as well as the terms appearing in fewer than 10 tweets. Second, based on the document, the LDA will generate the topic distribution, which shows how many percent each topic take up in all the topics (Here we set the number of topics \( N = 100 \)). Necessarily, the sum of percentage of all the topics in a document equals to 100\%. Based on the topic distribution, we can define the topic entropy \( T_i \) of user \( i \) as below:

\[
T_i = \frac{1}{\ln N} \sum_{k=1}^{N} c_{ik} \ln \frac{c_{ik}}{c_i}.
\]

where \( c_{ik} \) refers to the percentage of the topic \( k \) for the user \( i \), \( c_i \) is the sum of the percentages of all the topics of the user \( j \) (here \( c_i \) always is 1), and \( N \) is the number of the topics, 100. For a user, if its tweets only focus on one topic, its topic entropy will be 0; while if its tweets cover all the 100 topics uniformly, its topic entropy will be 1. Hence, higher entropy denotes wider range of topics in the tweets of users.

D. The summary of Features

We summarize all the features we proposed for the prediction in Table II. The table presents the abbreviation name and the description of each feature. Next we will show whether these features are significant for predicting the user influence and how the performances of the predictive models will be.

V. REGRESSION ANALYSES AND PREDICTION

Here we first present the logistic regression analysis to show the correlation between the features we proposed and users influence, and then we conduct the predictions via the four models: Support Vector Machine, J48 decision tree, Naive Bayes, and Bagging.

A. Experiment Setup

We further filter users in Section III for the experiments, because some of them lack corresponding data for computation of certain features. For example, because some users have no information of time zones in their profiles, we cannot transfer the published time of tweets into the local time of users to identify the period of day time or night time, and further cannot compute the feature of the time entropy. We filter these kinds of users, and as a result, 11,025 users are selected for the experiments.

For these selected users, their influence will be measured by the average click number per day received by their URLs. Rather than predicting the exact click number, we define several categories to represent the levels of the influence and predict which categories the users will belong to. The reason is because these categories can provide a clear concept about the levels of user influence and this method can compute the predictive accuracy which is more apparent to describe the performance of the prediction. Specifically, we divide the users into four categories depending on the average click number per day, and these classification results are shown in Table III.

<table>
<thead>
<tr>
<th>Table III.</th>
<th>THE CATEGORIES OF USERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Range of click number</td>
</tr>
<tr>
<td>1</td>
<td>0 - 10</td>
</tr>
<tr>
<td>2</td>
<td>10 - 100</td>
</tr>
<tr>
<td>3</td>
<td>100 - 1000</td>
</tr>
<tr>
<td>4</td>
<td>1000 +</td>
</tr>
</tbody>
</table>

B. Regression Analysis

Before conducting the prediction, we first explore the correlation between the features and the influence, i.e., whether the features we proposed have a predictive power for predicting user influence and how significant the features are for the prediction. To this end, we employ logistic regression analysis when using all the features to predict the categories of the influence. Table IV presents the results of the regression analysis.

Apparently most of the features we proposed have a significant predictive power to the user influence with
levels of significance of less than 0.0001, except for the three features: domain, mention and mentionHashtag. These three features have high significant levels, i.e., they have week relationship with the user influence. It should be noted that the positive or negative of coefficient in the regression might not indicate the positive or negative correlation between the features (independent variables) and the categories of user influence (dependent variable), because some features have high correlation. For example, the number of followers has a negative coefficient in the regression analysis because it is highly correlated with the number of active followers. Indeed, when we remove the number of active followers from the regression, the coefficient of the number of followers becomes positive and this feature still remains significant.

The results of the regression analysis thus provide strong evidences that most of the characters from user properties, behavior and topics affect the user influence. The features we defined, such as the day entropy, user topic category and topic entropy, have a significant predictive power to the influence.

C. Influence Prediction

To identify the better model, we select four widely used methods: Support Vector Machine (SVM) classification, J48 decision tree, Naive Bayes, and Bagging. For the model of SVM, we use the LIBSVM [24], which is an integrated software that implemented the SVM to conduct classification. And for this model, we use the popular e-SVR algorithm with a kernel function of Radial Basis Function (RBF). For the last three models, we use the Weka [25], a collection of machine learning algorithms for data mining tasks, to perform experiments. All methods are employed with 10-fold cross-validation. And their performances are evaluated by comparing the accuracy and F-score. The accuracy is the proportion of true results in the population. Assume that $tp$ are true positive, $fp$ - false positive, $fn$ - false negative, and $tn$ - true negative counts, and the accuracy can be computed as below:

$$\text{accuracy} = \frac{tp + tn}{tp + fp + fn + tn}.$$  \hspace{1cm} (6) The F-score combines Recall and Precision with an equal weight, in the following form:

$$F - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$  
$$\text{Recall} = \frac{tp}{tp + fn}.$$  \hspace{1cm} (7) \hspace{1cm} (9) where Precision represents the proportion of the true positives against all the positive results and Recall shows the proportion of the true positives against the positive and false negative results. And both they are computed as below:

$$\text{Precision} = \frac{tp}{tp + fp}.$$  \hspace{1cm} (8)

The results are shown in Table V. We can see that the methods have a little impact on the predictive performance. For example, the difference between the best accuracy and worst one is around 6%. The Bagging model achieves the best performance in both accuracy and F-score, and the overall accuracy arrives at more than 82% in determining whether a user will belong to a low-influence, medium-influence, high-influence, or extra-high-influence group.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
<th>F-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>77.10</td>
<td>74.16</td>
</tr>
<tr>
<td>J48 Decision Trees</td>
<td>80.61</td>
<td>80.01</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>75.99</td>
<td>76.86</td>
</tr>
<tr>
<td>Bagging</td>
<td>82.55</td>
<td>81.96</td>
</tr>
</tbody>
</table>

The predictive results are shown in Table V. We can see that the accuracy is around 82% in determining whether a user will belong to a low-influence, medium-influence, high-influence, or extra-high-influence group.

VI. CONCLUSIONS

In this paper, we predicted the user influence based on the standard of the accurate click number of URLs. We first exploited a wide range of possible features consisting of the sets of user properties, behavior and topics. These features not only include the basic properties, such as the number of followers, friends and lists, but also include our defined features, such as the entropies of published time and topics. And then we defined four categories based on the click number to represent the levels of user influence. After that we conducted the logistic regression analysis to identify whether the features have a predictive power to predict user influence, and find that most of the features, such as the number of followers, number of tweets, time entropy, topic category and entropy have a significantly predictive power. Finally, by using four models: SVM, J48 Decision Trees, Naive Bayes and Bagging, we predicted the levels of user influence, and find that the models have a little impact on the predictive performance and the Bagging model achieve the best result with an overall accuracy of more than 82% in determining whether a user will belong to a low-influence, medium-influence, high-influence, or extra-high-influence group.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation of China (NSFC), Grant No. 61272527.
REFERENCES