Prediction algorithm for users Retweet Times

Haihao Yu, Xu Feng Bai, ChengZhe Huang, Haoliang Qi

Heilongjiang Institute of Technology, Harbin, China

Abstract. In view of the fact that the propagation path topology cannot effectively deal with complex social network consists of hundreds of millions of users. More researchers choose to use machine learning methods to complete retweet prediction. Those use the classification method to judge whether a message will be retweeted or not. This paper argues that retweet prediction should be regression analysis problem, not just the classification problem. Through collecting user characteristics on Twitter and selecting some features which have an important impact on the retweet behavior, a Prediction algorithm Based on the Logistic Regression for users Retweet Times in social network was proposed. Experiment results based on the actual data set show the regression analysis predicting model has a good predicting accuracy in dealing with retweet predicting, the proposed method is effectiveness.

Keywords: Social network, Retweet Times, tweet, Logistic Regression

1 Introduction

With the development of internet and mobile technology, In particular Microblog has greatly accelerated the speed of information dissemination in the network. Using microblog Retweet, users can very easily share Microblog content of other users to achieve of information dissemination. A microblog after forwarding of different users, the propagation speed will increase in a geometric level. The objective of retweet predicting is to predict accurately a message transmission range and development trends. Through this information, the process of information dissemination can be effective intervention to control the spread of the message [1]. This study is very important for many fields, such as viral marketing, personalized message recommendation [2] and others.

Currently, the research of retweet prediction on social network is divided into two directions. One is through studying the information propagation path topology to build prediction model, most research based on dynamic propagation and virus propagation theory [4]. The other way to build prediction model is based on machine learning algorithm, consists mainly of Support Vector Machine [5-7].

The research of information propagation path topology is base on determining a node whether forwarding. With large user bases and complex user relation in The whole network topology, it's a very difficult task that construct the topology of user networks

In view of the fact that the propagation path topology cannot effectively deal with complex social network consists of hundreds of millions of users. More researchers
choose to use machine learning methods to complete retweet prediction. Those use the classification method to judge whether a message will be retweeted or not in the future. However, they only studied whether the message will be retweeted without taking into account the retweet times of the message.

In summary, these previous results are unsatisfactory, further studies are still necessary. This paper argues that retweet predicting should be regression analysis problem, not just the classification problem. Because we need through the forwarding times of the message to determine the propagation scale, rather than whether retweeted. This paper reposts on building regression analysis model based on Logistic Regression algorithm to predicting the scale of information dissemination. We selected some features that have an important impact on the retweet behavior and divided into four categories, including user features, text features, temporal feature and metadata feature. To note is that we take into account the effect of text content of the retweet behavior.

2 Retweet predicting based on regression analysis

Regression analysis. Regression analysis is a statistical method which is handling the correlation between variables. By one or more value of the variable, predict the value of another variable. Further predict the scale of development and factor analysis.

Retweet Feature selection. Because logistic regression generally used to solve linear regression problem, Based on the analysis of the relevant research, this paper choose microblog features of linear relationship with the forwarding and greater impact. We use the following features and divide them into for distinct sets: User features, Text features, Temporal feature and Metadata feature.

Logistic Regression model. In statistics, Logistic Regression is a statistical classification model, used to predict the results of classification based on one or more of the features. Logistic Regression measures the relationship between variable independent variables and categorical dependent, by using probability scores as the predicted values of the dependent variable. Thus, Logistic Regression not only be used for solving classification problems also be used to solve the regression problem, that is why we use Logistic Regression modeling.

Logistic Regression (LR) and Support Vector Machine (SVM) are discriminative learning model. They can be used to establish the prediction model. But Support Vector Machine use quadratic programming for support vector. Algorithms for quadratic programming involve calculation of m-order matrix. (m is the number of samples). When the m is large, the storage and computation of the matrix will consume a large amount of machine memory and CPU. Space complexity and Time Complexity of LR is lower than SVM.

This paper uses multiple microblog characteristics during establishing the prediction model based on LR. Therefore, the model has a number of independent variables. The logistic function can be written as:
\[
\log \text{it} (p_j) = \ln(\frac{p_j}{1-p_j}) = a_j + b_1 x_1 + b_2 x_2 + \ldots + b_q x_q
\]

Where, \(x_1, x_2, \ldots, x_p\) is independent variable. That is microblog characteristics.

\(p_j = p(y \leq j | x)\) is dependent variable (Retweet Times), \(y\) is cumulative probability of \(j\) independent variables.

### 3 Experiments

**Datasets.** We run our experiments with tweets collected in February and March 2013 which obtained from Twitter API by ourselves, because need to analyze the topic of each tweet, we retain only the English text tweets. After a few preliminary treatments, the final data set contains approximately 136.8 million tweets and 24.56 million users.

**Table 2.** Detailed information of each dataset

<table>
<thead>
<tr>
<th>dataset</th>
<th>number of tweets</th>
<th>retweet rate</th>
<th>maximum number of forwarding times</th>
<th>average number of forwarding times</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset one: training set</td>
<td>30043708</td>
<td>7.813%</td>
<td>2510</td>
<td>2.163</td>
</tr>
<tr>
<td>dataset one: testing set</td>
<td>30539275</td>
<td>8.026%</td>
<td>2444</td>
<td>1.977</td>
</tr>
<tr>
<td>dataset two: training set</td>
<td>30254902</td>
<td>8.244%</td>
<td>2476</td>
<td>1.901</td>
</tr>
<tr>
<td>dataset two: testing set</td>
<td>30530744</td>
<td>8.351%</td>
<td>2464</td>
<td>1.985</td>
</tr>
<tr>
<td>dataset three: training set</td>
<td>30349052</td>
<td>7.923%</td>
<td>2553</td>
<td>2.082</td>
</tr>
<tr>
<td>dataset three: testing set</td>
<td>30067348</td>
<td>8.171%</td>
<td>2507</td>
<td>1.994</td>
</tr>
<tr>
<td>dataset four: training set</td>
<td>30192749</td>
<td>8.393%</td>
<td>2496</td>
<td>1.920</td>
</tr>
<tr>
<td>dataset four: testing set</td>
<td>30412894</td>
<td>8.125%</td>
<td>2466</td>
<td>1.969</td>
</tr>
</tbody>
</table>

By definition of forwarding times, the time span of training set and test set need to be consistent, furthermore, in order to reduce the influence of data set on the results, the data set is divided into four parts. Thus, each training set or test set contains a week of tweets, and to ensure that the maximum forwarding times of training set is greater than the test set. As shown in Table 2, we counted the number of tweets, retweet rate and the maximum number of forwarding times for each part. We can also find that each part contains about 30 million tweets and there are only 8.393\% of the tweets has been reweeted in our datasets.
**Evaluation standards.** We use MSE (Mean Square Error) to evaluate the prediction results of different models, the smaller value of MSE, the better prediction results of model, and the final results should be the average value of four experimental, function can be written as:

\[
MSE (\hat{y}, \bar{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2
\]

Where, \( \hat{y}_i \) is the prediction value of sample \( i \), \( y_i \) is the calculated value of sample \( i \) According to the Retweet Times definition above.

**Experimental settings.** We build classification model Using the Hong[8] and regression model above by each part of dataset, In addition, we use MSE to evaluate the prediction results of different models, the smaller value of MSE, the better prediction results of model, and the final results should be the average value of four experimental.

For Regression model, we establish Logistic Regression model using training set, and in the process of model building we use four features: User features, Text features, Temporal feature and Metadata feature. Then prediction retweet Times according to the established prediction model. Finally calculate the mean square error of prediction value and the real value for tweets.

For classification model, instead of directly predicting the forwarding times, we divide the messages into different classes by the forwarding times: a: zero, b: between 1 and 100, c: between 101 and 1000, d: more than 1000. When calculating the MSE value, we need to give a value for each category as their predicted forwarding times: a: 0, b: 50, c: 550 and d: 1000. For classification model and calculate the mean square error of prediction value and the real value for tweets.

**Experimental results and analysis.** The detailed information of each dataset are shown in Table 1, we can see that the max of forwarding time only 2553, this is not consistent with the real situation. Getting all tweets which user generated is an impossible task, because Twitter did not fully open the data interface, we can only crawl the data randomly in a time sequence from Twitter API. Thus, the experimental data will be biased with the actual situation, but it will not affect the experimental results, to some extent, this will reduce the advantage of the regression model.

**Table 2. The experimental results**

<table>
<thead>
<tr>
<th></th>
<th>classification model(MSE)</th>
<th>regression model(MSE)</th>
<th>performance improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>dataset one</td>
<td>315.145</td>
<td>219.542</td>
<td>30.34%</td>
</tr>
<tr>
<td>dataset two</td>
<td>384.362</td>
<td>299.138</td>
<td>22.17%</td>
</tr>
<tr>
<td>dataset three</td>
<td>446.320</td>
<td>342.764</td>
<td>23.20%</td>
</tr>
<tr>
<td>dataset four</td>
<td>426.735</td>
<td>338.258</td>
<td>20.73%</td>
</tr>
<tr>
<td>mean</td>
<td>393.141</td>
<td>299.93</td>
<td>23.71%</td>
</tr>
</tbody>
</table>

The data in Table 2 also show that the MSE of regression model based on Logistic Regression is smaller than classification model, the average of the former reached 393.141, the latter only 299.93, the prediction performance of the model is improve
about 24%. In addition, there are 92% tweets of the dataset have not been retweeted, that their forwarding times is 0, both regression model and classification model will predict the forwarding times is 0 if the prediction is correct, then will narrow the gap of MSE between regression model and classification model. Despite this, the predicted performance of regression model still better than the classification model.

4 Conclusion

This study predicts the forwarding times of a tweet through building a regression model based on Logistic Regression algorithm. We selected some features which have an important impact on the retweet behavior and build the regression analysis predicting model. Experiment results based on the actual data set show the regression analysis predicting model has a better predicting accuracy in dealing with retweet predicting than the classification model, our method is effectiveness.

References

5. Petrovic S, Osborne M, Lavrenko V. RT to Win! Predicting Message Propagation in Twitter[C]//The International AAAI Conference on Weblogs and Social Media (ICWSM), (2011)