International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.5 (2015), pp.315-322
http://dx.doi.org/10.14257/ijmue.2015.10.5.29

Prediction of Users Retweet Times in Social Network

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

Heilongjiang Institute of Technology, Harbin, China
haoliangqi163@163.com;hqgsyhh@163.com

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 constructs 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.

ISSN: 1975-0080 IJMUE

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. Related Work

With the sweeping of social network, the associated research achieves widespread concerns from academia to business sector. How to analyze and study the scale of information dissemination has become an important research direction. Researchers made some research results in retweet predicting from different perspective.

Based on Propagation Topology: To building prediction model by information propagation path topology, the topology should include all the users who have seen the tweet, then through analyzing the similarity between users and user history behavior to decide whether the user will retweet or not. Macskassy [9] build four prediction models including recent communication model and homophily model, and got best prediction performance when using all models. Yantao [10] also proposed a new algorithm to establish the information propagation path topology. The limitations of their research is that their model is only suitable for small scale users, however, the amount of users in social network is so huge that cannot build prediction model.

Based on Machine Learning: Because there are enormous difficulties to build prediction model by information propagation path topology, the other scholars choose to use machine learning algorithm to build the prediction model. Sasa [8] used the PA algorithm in the study. The author also makes a statistical experiment by artificial prediction. The experimental results show that the result of prediction model of machine learning algorithm is superior to the artificial prediction. And Hong [12] used Logistic Regression, they successfully build the prediction model and enhanced the user scale, the number of users was increased to 2.5 million by Hong. However, the research only studied whether the message will be retweeted without taking into account the retweet times of the message will be received. The study is still qualitative research as a classification problem.

Retweet Times Definition: Retweet mechanism of different social media are not all the same, different retweet mechanism caused by different of retweet times definition. For example, in Twitter only the original microblog can be retweeted, Forwarding microblog does not obtain forwarding opportunities. However, In Sina microblog forwarding microblog can also obtain forwarding opportunities. Therefore, before the establishment of regression model, we need to give a clear definition of Retweet times. This paper adopts the definition of Retweet times in the literature [12]. First, we give each micrblog text an identifier, Text content identical microblog get the same identifier, Then all microblog were classified according to those identifier. Two messages are considered identical if they share the same identifier. We sort all such messages by ascending time order, forming a chain of messages.

Whether definition retweeted: for n messages in a chain, n-1 messages are considered as "positive instances" and the last one as a "negative instance". Where $\mathrm{n} \geq 1$.

Retweet Times definition: For a message as "v_i" in a chain, Retweet times are considered as $n-i$. Where $\mathrm{i} = 1,2,\ldots,\mathrm{n}$

3. Retweet Predicting based on Regression Analysis

Information Serialization

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 constructs the topology of user networks. As Figure 1:

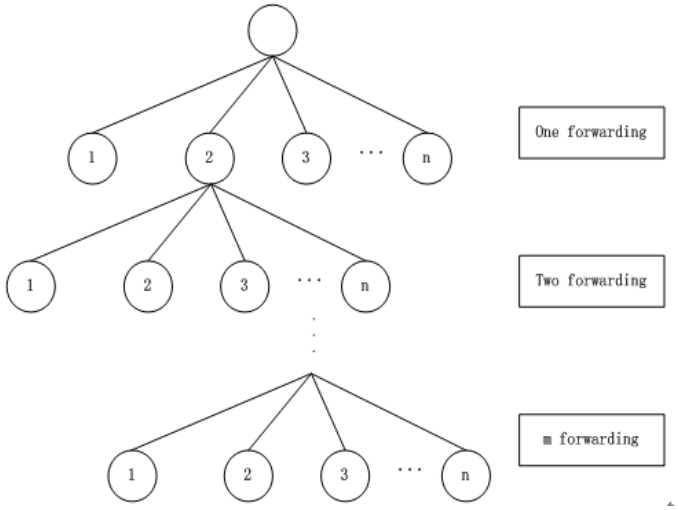

Figure 1. Information Propagation Path Diagram

Figure 1 Describe an original microblog propagation, define n as user's fans number, m as retweet times, k as user's node number, then $K=n^m$,if fans number n is big and retweet times m is more, then user's node number is very large, for example ,if Katy Perry(@katyperry, 59700000 fans), Taylor Swift(@taylorswift13, 46500000 fans) and Lady Gaga(@ladygaga, 427000000 fans) are involved in a microblog retweet, the user's number and complex user relation of this microblog Will lead to the space complexity and Time Complexity is high for establishment of complete information propagation path topology, Therefore, the method is not appropriate.

We used serial processing mode in the user node analysis, as Figure 2, Through the serialization, we avoid the complicated relationship between the users, reduce the workload and the space complexity and Time Complexity.

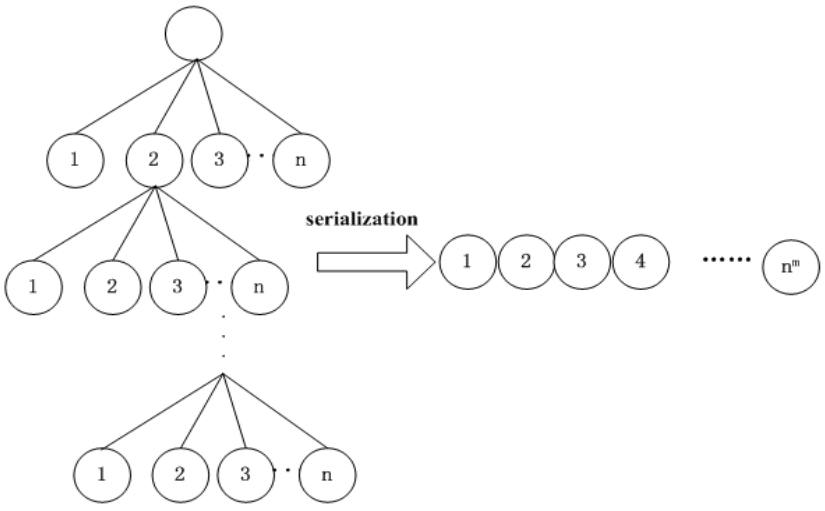

Figure 2. Information Serialization Diagram

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: during establishing the prediction model based on LR, In order to obtain more accurate prediction, we will select significant microblog characteristics. Before this paper, many researchers have conducted studies retweet feature selection. The largest use of microblog feature is the number of fans and attention. The feature of the text Whether including URL,"#"(topic), "@" (mention others)or not also was studies on retweet feature selection.

In Sasa [8] study, Whether the Text including URL,"#" or not, The number of fans,The number of attention have a positive impact on the retweet. but whether mention others and the number of historical statuses have negative effects. In addition, user registration time has little effect on the retweet At the same time, the study found 91% of the be forwarded microblog is released by the authenticated user. Sasa[8] found the microblog at different time in a day have different features. For example, the released microblog at nine o'clock is longer. Released Microblog amount is the largest At Thirteen o'clock. microblog between 10 o'clock and 12o'clock pay more attention. Artzi [11] Found Whether forwarding microblog and text including other media information also influence on the forwarding behavior.

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, the specific feature as the Table 1:

Table 1. Microblog Feature Classification

Feature type	Feature name
User features	number of followers, number of fans, the user was listed is the user verified number of historical statuses
Text features	Whether including hashtags Whether mention others Whether including url Whether including media information
Temporal feature	the time period when the information release
Metadata feature	Whether Forwarding microblog

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:

$$logit(p_j) = \ln(\frac{p_j}{1-p_j})$$
$$= a_j + b_1 x_1 + b_2 x_2 + ... + b_q x_q$$

Where, $x_1, x_2 ... x_p$ is independent variables. That is microblog characteristics.

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

4. 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

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

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:

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

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

Experimental Settings: We build classification model Using the Hong[12] 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 3. The Experimental Results

	classification model(MSE)	regression model(MSE)	performance improvement
dataset one	315.145	219.542	30.34%
dataset two	384.362	299.138	22.17%
dataset three	446.320	342.764	23.20%
dataset four	426.735	338.258	20.73%
mean	393,141	299.93	23.71%

The data in Table 3 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.

Table . Contrast Experiment of Predicted Value and the True Value

ID	Real value	Classification	Regression
1	0	0	3
2	0	0	0
3	0	50	4
4	0	0	2
5	6	50	30
6	24	0	25
7	30	50	22
8	49	50	37
9	61	0	3
10	88	50	67
11	42	50	30
12	177	550	315
13	231	50	108
14	647	50	683
15	1354	0	767
16	1566	550	923
17	2179	50	858

In this paper ,we randomly selected a number of different orders microblog According to the actual forwarding number, The three values(the actual value, classification model predicted values ,d regression model prediction values) are compare as Table 4, From table 4 we can see, For the microblog forwarding times Equal to 0 , The predicted results of classification model is closer to the real value, The predicted results of the regression model in the small fluctuation, But the error is very small. For the microblog forwarding times below 100, the regression model and classification model have better prediction precision. For the microblog the forwarding times larger, the predicted results of regression model is much better than the classification model. This shows that regression model has good prediction accuracy for prediction algorithm of users retweet times, the algorithm can be used in practical application.

5. 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.

.

Acknowledgements

This paper is supported by Heilongjiang Province Educational Committee Science Foundation (12541683).

References

[1] J. Xie and G.-s. Liu, “Prediction of User’s Retweet Behavior in Social NetWork”, Journal of SHANGHAI JIAO TONG University, vol. 47, no. 004, (**2013**), pp. 84-588.

[2] I. Konstas, V. Stathopoulos and JM. Jose, "On Social Networks and Collaborative recommendation [C]", Proceedings of the 32th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, (**2009**), pp. 195-202.
[3] Najar, L. Denoyer and P. Galllinari, "Predicting information diffusion on social networks with partial knowledge [C]", Proceedings of the 21st international conference companion on World Wide Web, ACM, (**2012**), pp. 1197-1204.
[4] M. Gomez Rodriguez, J. Leskovec and A. Krause, "Inferring networks of diffusion and influence [C]", Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, (**2010**), pp. 1019-1028.
[5] S. Petrovic, M. Osborne and V. Lavrenko, "RT to Win! Predicting Message Propagation in Twitter[C]", The International AAAI Conference on Weblogs and Social Media (ICWSM), (**2011**).
[6] TR. Zaman, R. Herbrich and G. Van, "Predicting information spreading in twitter [C]", Workshop on Computational Social Science and the Wisdom of Crowds, NIPS, vol. 104, no. 45, (**2010**), pp. 17599-601.
[7] Z. Luo and M. Osborne, "Who Will Retweet Me? Finding Retweeters in Twitter [C]", Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, (**2013),** pp. 978-981.
[8] P. Sasa, O. Miles and L. Victor, "RT to Win! Predicting Message Propagation in Twitter[C]", The International AAAI Conference on Weblogs and Social Media (ICWSM), (**2011**).
[9] SA. Macskassy and M. Michelson, "Why Do People Retweet Anti-Homophily Wins the Day! [C]", The International AAAI Conference on Weblogs and Social Media (ICWSM), (**2011),** pp. 209-216.
[10] J. Yantao, W. Yuanzhuo and L. Jingyuan, "Structural-interaction link prediction in microblogs[C]", Proceedings of the 22nd international conference on World Wide Web companion，International World Wide Web Conferences Steering Committee, (**2013**), pp. 193-194.
[11] Y. Artzi, P. Pantel and M. Gamon, "Predicting Responses to Microblog Posts[C]", 2012 Conference of the North American Chapter of the Association for Computational Linguistics, (**2012**), pp. 602-606.
[12] L. G. Hong, O. Dan and B. D. Davison, "Predicting Popular Messages in Twitter", Proceedings of the 20th international conference companion on World Wide Web, ACM, (**2011**).
[13] L. Robert and R. Deriche, "Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities", Proceedings of the 4th European Conference on Computer Vision. (1996) April 15-18; Cambridge, UK.

Authors

Haihao Yu, He was born in Huanan, China in 1974. He received her Bachelor of Engineering degree in Welding Technology and Equipment from Jia MuSi University in 1997 and the Master of Engineering degree in Technology of Computer Application from Northeast Forestry University in 2008. Her major field of study is Artificial Intelligence.

Haoliang Qi, he born in 1972, Ph.D., professor, Longjiang Scholor (Distinguish Professor of Heilongjiang Province). His current research interests include natural language processing, information retrieval and Web information processing. ***Corresponding author**
Tel: 13100875759; E-mail: haoliangqi163@163.com.

