International Journal of u- and e- Service, Science and Technology
Vol.9, No. 9 (2016), pp.373-382
http://dx.doi.org/10.14257/ijunesst.2016.9.9.34

Prediction Based on the Acquisition of Sina Weibo Data and Users' Forwarding Behavior

Lufei Wang

Beijing Jiaotong University, Beijing, China
13121688@bjtu.edu.cn

Abstract

Microblogging, also known as Weibo, has currently become the prevalent social networking platform, and its function of disseminating information has been increasingly recognized and valued. One of the main approaches of information communication on Weibo is forwarding activities among users. Therefore, the key interests have focused on the microblogs that are easily recognized and forwarded. Through predicting the possibility of a micro-blog being forwarded among users, it is able to effectively improve the efficiency of information spreading on the Weibo platform. On the basis of Sina Weibo data, this paper intends to study 13 features of Sina Weibo including user and content, so as to conduct machine learning, to establish a feature analysis model, to identify key influence factors on Weibo forwarding, to study the combination of feature attributes for the first time, to explore the correlation among features, then to predict the probability of a micro-blog being forwarded by users. Meanwhile, concerning different features held by different data, different machine learning algorithms are applied in model training, in order to identify the algorithm with the highest prediction accuracy.

Keywords*: Weibo, forwarding behavior, frequency of forwarding, machine learning, feature extraction, prediction*

1. Introduction

With the continuous development of computer technology, especially WEB2.0 technology, the Internet has been gradually affecting the way people live. The emergence of Facebook [1], Twitter [2] and Youtube [3] marks the arrival of a new pattern of the Internet. Based on these three social networks [4], the way information transmits among users also has experienced a number of changes and increasing diversity.

The emergence and the rise of social networks are unwittingly changing people's life styles, social behavior and the mode of information access. As one of the major platform of social networks, Weibo has attracted increasing attention. In recent years, Weibo has seen rapidly development, and become an important platform for information dissemination. Weibo is featured by strong interactivity, numerous participation, wide coverage, fast spread, low overhead and other advantages. It is of great significance to analyze the principle and mechanism of online information dissemination on Weibo. Researchers could analyze the reason for information communication, Weibo users' behavior, hobbies and interests and so on, thereby figuring out issues occurring in the human society, such as group activities, hobbies and trends. In addition, companies could understand behaviors, interests and hobbies of different types of Weibo users, so as to deliver a more precise advertisement targeting.

Fundamentally, Weibo is still a communicative media, with an ultimate goal of passing messages onto the outside world and gaining maximum effects of communication. Compared with traditional media, this newly emerging social media holds a number of unique features. Thus, it is of particularly necessity to study how to employ Weibo to effectively and efficiently communicate information in the new media environment.

ISSN: 2005-4246 IJUNESST

Communication effect is one concept in communication studies, which refers to the overall impacts and results of communication activities, especially activities of mass communicative media like the press, radio and television, made on the receivers and the society [5]. Communication effect is an abstract and qualitative concept. There exists currently no generally accepted standard to evaluate the communication effect. Different medias adopt different indicators to assess its effect of communication, such as newspaper circulation, television ratings, movie box office and so on. Weibo achieves sustained information dissemination through forwarding, and the number of forwarding could be regarded as a key indicator of communication effect. Therefore, the analysis of users' forwarding behavior is an important approach to predict the scope of forwarding [6]. Forwarding is the primary mechanism for information dissemination on the Weibo network. Different influence factors of Weibo users' forwarding behavior would lead to different forwarding behavior of users with different types of link relations.

This paper intends to extract 13 features attributes from the available Weibo data, including users' basic information, natural language processing of Weibo content, and a simple emotion analysis. The third chapter will elaborate on the specific feature attributes. Then, SVM algorithm is used to conduct training process, and 13 features attributes are applied to establish a prediction model to predict the possibility of Weibo forwarding. Meanwhile the attributes that mostly affecting prediction accuracy will be identified. On this basis, combinations of feature attributes that affect the final results will be further studied. The fourth chapter includes the analysis of experimental results, a summary of the study and plans for future research.

2. Literature Review

More attention from domestic and foreign researchers has been attached to one highlighted issue: the analysis on social networks users' behavior and information dissemination. Twitter appears earlier than Weibo in China, with a more extensive users in Europe and America. Thus, the study of foreign researchers mainly focuses on various aspects of Twitter.

Haewoon Kwak [6] *et. al.*, first introduce what Twitter is: Twitter is an online social networking service that allows users to send and read short messages within 140 characters, and the message is known as "tweet". Danah boyd *et. al.*, [10] select Twitter as the main object of study. The definition and the emerging background of Twitter are first introduced. A detailed introduction to retweeting behavior (similar to the function of forwarding in Weibo) is made, including how to forward, why to forward, and which tweets to forward.

Wolfram *et. al.*, [8] apply the SVM model to directly extract features from microblogging text for model training and to predict the Nasdaq index. Han Yideng *et. al.*, [9] construct a time series model on the trends of microblogging topics, and make predictions through K-nearest neighbor (KNN). Yao Haibo *et. al.*, [10] build up a multiple regression model to predict the trends of topics, based on features like participation rate of opinion leaders, forwarding rate of microblogging, and comment rate of microblogging.

Yang Zhang, and Zhiheng Xu *et. al.*, [11] use information gain (IG) to set different influence made by different twitter features on forwarding behavior as different weights. The higher the weight of one feature is, the larger the impact on forwarding is. By comparing two classification algorithms—SVM and logistic regression, it is indicated that weight model produces more satisfactory effects in predicting users' forwarding behavior. However, this method holds two disadvantages. One is that the extracted features do not involve tweet content, such as considering discussion topics and emotions analysis; the other is to leave out the prior correlation among these features. Bongwon Suh *et. al.*, [12] employ principal component analysis (PCA) to investigate the major features that are most influential to users' forwarding behavior. A prediction model is constructed by

combining these features and the generalized linear modeling (GLM), which enables a discussion about the relations between user features and its forwarding behavior. The conclusion is that, if a tweet contains URL and hashtag, then this Tweet is more likely to be forwarded. In terms of user feature, the number of followees, the number of followers, and the length of registration also have a great impact on the possibility of being forwarded. However, research of this kind limit itself to statistical analysis on users, and it does not take advantage of these user features to further predict forwarding behavior. Yang *et. al.*, [13] analyze the impact of retweeting behavior made on users, tweets content and time, and they also predict retweeting with a semi-supervised graph model algorithm. Li Ying Yue *et. al.*, [14] develop the forwarding tree through the microblogging forwarding path, and predict each forwarding behavior on the forwarding path using iterative method.

By extracting relevant features, Zhang Shengbing [15] and other researchers have found out that features like homogeneity difference, micro-network structure, geographical distance, and gender exert certain impacts on microblogging forwarding behavior. Among which, homogeneity difference is the most influential factor. Luo Zhilin *et. al.*, [16] extract microblogging features like weight ratio, users' personal information to study users' forwarding behavior. Then the prediction algorithm based on random forest (RFMR) is adopted to forecast forwarding behaviors. Danah boyd *et. al.*, [17] employ Twitter as the main object of study. The definition and the emerging background of Twitter are first introduced. Then retweeting behavior is detailedly introduced, including how to forward, why to forward, and which tweets to forward. Yoav Artzi *et. al.*, [18] divide twitter data into six features: history feature, social feature, aggregate lexical feature, local content feature, posting feature, sentiment feature. The method of multiple additive regression trees (MART) is used to train data, and one of six above features is removed one by one. The feature with the highest prediction accuracy is identified, which is proven to exert the largest impacts on retweeting.

Zhuchenluo *et. al.*, [19] pursue another approach and study on which users will retweet from the perspective of forwarders. By focusing on the relationship among followers, they analyze six features including retweet history, follower status, follower active time and follower interests, and further find out the user who is most likely to forward a microblog. Lee *et. al.*, [20] study forwarding behavior among strangers on Twitter, in order to predict the probability of users' forwarding when a stranger mentions (@) one user and requests to be forwarded. A recommendation system is established, and users who communicate information are more inclined to be recommended.

The above study involves a variety of machine learning algorithms and prediction models, and different features are extracted and trained. However, the correlation among these features are not properly analyzed. Therefore, this paper proposes the impact of combined features on the prediction of microblogging forwarding, and the degree of correlation among various features is analyzed through a few experiments.

3. Study Content

This chapter aims to extract 13 feature attributes through the acquired Weibo data, including users' basic information, natural language processing of Weibo content, and a simple sentiment analysis. SVM algorithm is adopted for training process, and a prediction model is established to realize a possibility prediction of microblogging forwarding. At the same time the attributes that mostly affecting prediction accuracy will be identified. On this basis, a combination of feature attributes that affect the final results will be further studied. We will analyze which algorithms are suitable for Weibo messages with various features, thereby further improving the prediction results.

3.1. Data Processing

Data source in this paper includes over 800,000 pieces of Weibo data with the help of web crawler, which are saved in the Mysql database. From the perspective of user features and microblogging features, a total of 13 feature attributes are extracted, as shown in the following Table 1.

Table 1. Weibo Features

No.	*Name*	*Comment*
1	Comment	The number of comments of this microblog from other users
2	Attitude	The number of "likes" of this microblog from other users
3	Gender	The sex of this user at registration
4	Follower	The number of followers
5	Followee	The number of followees
5	Status	The number of posted microblogs
7	Favorite	The number of microblogs favorited
8	URL	The number of microblogs that contain URL
9	Hashtag	The number of microblogs that contain "#"
10	Atnum	The number of microblogs that contain @
11	Positive_word	The number of microblogs that contain positive word
12	Negative_word	The number of microblogs that contain negative word
13	Province	The number of days of registration

First, the extraction of positive word and negative word uses the ICTCLAS system developed by the Institute of Computing Technology, Chinese Academy of Science to separate words. Then, the stop word list provided by China National Knowledge Infrastructure (CNKI) is applied to remove stop words. In the end, according to the word collection for sentiment analysis (beta version) [21] (ZHSD) released by CNKI, the Chinese sentiment dictionary constructed by National Taiwan University (NTUSD), and Chinese emotion word ontology (DSLD) [22], positive words and negative words contained in Weibo content are located. The statistical table of sentiment dictionary is shown in Table 2.

Table 2. Sentiment Dictionary

Sentiment dictionary	positive words	negative words	Total word count
ZHSD	936	1254	2090
NTUSD	2810	8276	11086
DSLD	13052	14322	27374
In total	16789	23852	40550

3.2 Evaluation Index of Text Classification

Performance evaluation index of text classification mainly comprises primarily of recall, precision, F1-score [50].

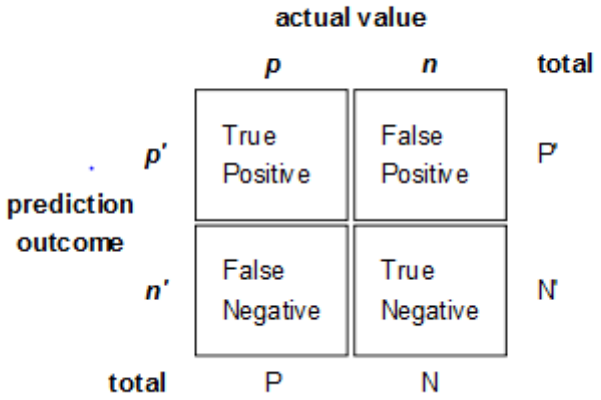

Figure 1. Confusion Matrix

In the confusion matrix, as shown in Figure 1, each instance could be classified as one of four types:

(1) True Positive, TP: Positive samples that are predicted as positive by the model

(2) False Positive, FP: Negative samples that are predicted as positive by the model

(3) False Negative, FN: Positive samples that are predicted as negative by the model

(4) True Negative, TN: Negative samples that are predicted as negative by the model

Then, evaluation indicators are calculated by the confusion matrix.

(1) Precision, P

Precision represents the percentage of the number of correctly classified positive samples and the actual number of classified samples in a category. Equation (3-3) is the computational formula for precision. Precision reflects the judgment capability of the classifier held for the whole sample, i.e. a positive result indicates positive, and negative result indicates negative. The higher the precision is, the smaller the error probability of the classifier in this category is.

$$P=TP/(TP+FP) \quad (1)$$

(2) Recall, R

Recall rate is also called true positive rate. It indicates that the percentage of the number of correctly classified samples and the number of duly classified samples in a category. Equation (3-4) is the computational formula for recall rate. Recall rate reflects the proportion of the number of correctly predicted positive samples in the total positive samples. The higher the recall rate is, the less the classifier might miss on this category.

$$R=TP/(TP+FN) \quad (2)$$

(3) F1 score

F1 score is an indicator used to measure the accuracy of dichotomous model. It is the harmonic mean of accuracy rate and recall rate, approximate to the smaller value between P and R. Equation (3-5) is the computational formula for F1 score, which could be seen as a weighted average of accuracy degree and recall rate of the model.

$$F=2*P*R/(P+R) \quad (3)$$

(4) Accuracy

Accuracy is the judgment capability of a classifier held for the whole sample, i.e. a positive result indicates positive, and negative result indicates negative. Equation (3-4) is the computational formula for accuracy rate.

$$A=(TP+TN)/(TP+FN+FP+TN) \quad (4)$$

3.2. Experiment Description and Results Analysis

Concerning the issues raised above, five experiments are conducted in this paper.

(1) Experiment 1:

First, 13 extracted features are formatted to work as a data set in this paper. The approach of 10 cross validation is adopted to generate training set and test set, which is then utilized to carry out two processes—no-weighted remodeling and weighted remodeling through SVM. Table 3-3, compares the predictions results of two models—weight and no weight. It indicates that the results of weighted model are evidently more favorable than those of no-weighted model. In Table 3-4, through genetic algorithms, weights are assigned to the values of 13 features, so as to identify a set of weights with the highest prediction accuracy. As demonstrated from the table, the number of comments, the number of @, and the number of likes make a significant influence on the final forwarding behavior.

Table 3. The Results of Prediction on Weight and No Weight

Model	Accuracy
Weight	0.9836
No weight	0.9789

Table 4. Feature Weights

No.	Name	Weight
1	Comment	0.1193
10	Atnum	0.1046
2	Attitude	0.0997
6	Status	0.0948
7	Favorite	0.0866
8	URL	0.0833
9	Hashtag	0.0817
4	Follower	0.0719
5	Followee	0.0686
11	Positive_word	0.0539

3	Gender	0.0523
12	Negative_word	0.0507
13	Account_days	0.0327

(2) Experiment 2:

Next, in order to verify whether the above weighted values assigned by 13 features through genetic algorithm correspond to reality, first, this paper resets the weighted values of 13 features to an equal value, in other words, the 13 features exert the same impact on the final prediction results. To find out which feature attribute has greater impacts on the possibility of being forwarded, the weight of each of the 13 features is reset to zero one by one. In other words, one feature attribute is regarded as being removed. Then, we follow the previous experiment procedure, and identify the feature attributes with greater impacts on prediction results. The experiment results are shown in Figure 2. The x-coordinate represents feature attribute, and the y-ordinate denotes the value of prediction accuracy after one feature is removed. When all features are retained and their weighted values are set as the same, the prediction accuracy is 97.89%. The red line indicates an average accuracy rate of 97.83% after 13 features are removed one by one. As seen from the figure, feature attribute 1, 10, 2, and 6 fall below this average value.

Feature attribute 1 is the number of comments on one microblog, which evidently affects the final prediction. It is indicated that, to some extent, the number of comments reflects the level of visibility of this microblog. The higher the visibility is, the greater the possibility of being forwarded by users is. Feature attribute 10 and 2 also have great impacts on the results of prediction, but smaller than that of feature attribute 1. In contrast, feature attribute 3—gender and feature attribute 7—the number of microblogs favorited by one user, have no significant impact on the final prediction.

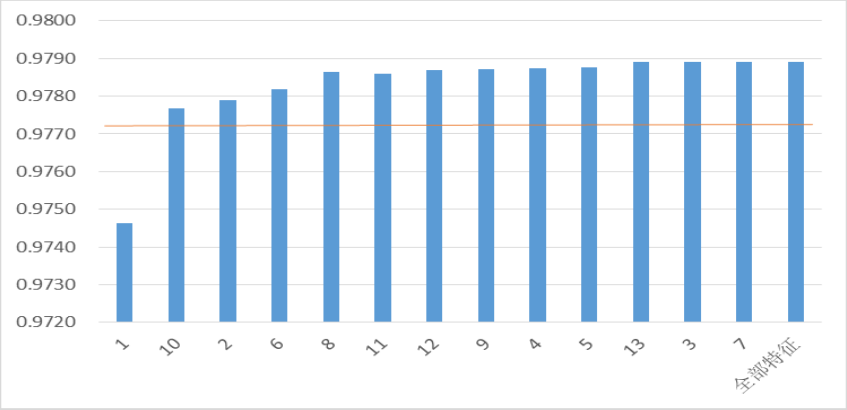

Figure 2. The Influence of Prediction Accuracy on Single Feature

(3) Experiment 3:

When feature attributes are removed one by one, some feature attributes have no obvious impact on prediction accuracy. However, it does not mean that these feature attributes do not affect microblogging forwarding. The possible reasons might be that certain feature attributes are more closely inter-related. In order to explore the fact, on the basis of above experiments, two feature attributes are removed in turn and microblogging forwarding is predicted. There are a total of 78 combinations, and the final average prediction accuracy is 97.71%. This paper selects combinations whose accuracy lies below this average value in the analysis. The experiment results are illustrated in Figure 3.

According to the figure, the successive removal of 13 feature attributes have a great impact on the final prediction. It is shown that feature attribute 1 is validly associated with

the other 12 ones, and the number of comments on one microblog significantly affects its forwarding. The more the comments are, the greater the visibility of this microblog has, which in turn means a greater probability of being forwarded by other applications. It is observed that the combination of feature attribute 4 and 6 has the largest influence on prediction accuracy. In previous experiments, when only feature attribute 6 is removed, the prediction accuracy is 97.82%, which falls under yet approximate the average value of 97.83%. So the impacts of feature attribute 6 on prediction results lie at the middle level. However, when both feature attribute 4 and feature attribute 6 are taken out, there occurs the most influential combination on prediction accuracy. It indicates that, if a user enjoys posting microblogs but has no followers, his microblogs are also difficult to be forwarded. But if the number of followers is larger, the possibility of his or her microblogs being forwarded is greater. In previous experiments, the removal of feature attribute 4 and feature attribute 5 alone produces little effect on the prediction results; however, when both feature attributes are taken out, the prediction accuracy falls far below the average, which indicates that this combination of largely influences the prediction results. The more followers users have, the more likely the microblogs are read by other users, and the higher the probability of this microblog being forwarded is. Another influential combination is feature attribute 5 and feature attribute 6. To be specific, the number of followees and the number of posted microblogs are closely related. One reason is that if a user has many followees, he or she has access to more microblogs from other users, and then is able to favorite or forward them, which increases the possibility of microblogging forwarding. The similar cases also include feature attribute 6 and 7, feature attribute 6 and 13, feature attribute 2 and 10, as well as feature attribute 2 and 6.

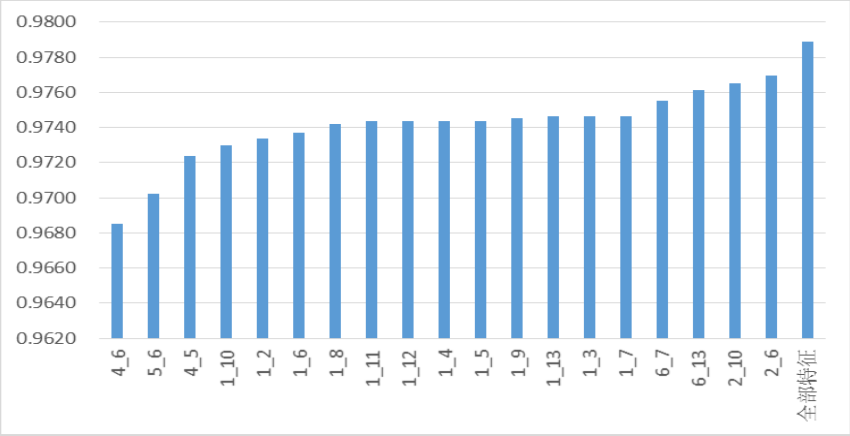

Figure 3. The Influence of Combination Features on Prediction Accuracy

(4) Experiment 4:

In Experiment 2, it is found out that feature attribute 1, 10, 2, and 6 are the four attributes that are most influential to prediction result, which are referred to as a feature set–A1. In Experiment 3, several highly correlated combinations of feature attributes are identified. Four more influential feature attributes—4, 5, 7, and 13 are added, which are referred to as A2. In this experiment, only A1, A2 as well as A1 and A2 are retained from the 13 features. Three experiments are conducted to compare the experimental results, as shown in Figure 4. As seen from the experimental results, eight feature attributes in A1 are the main factors affecting prediction results, which differs little from the prediction results of the 13 features. A1 is more influential than A2.

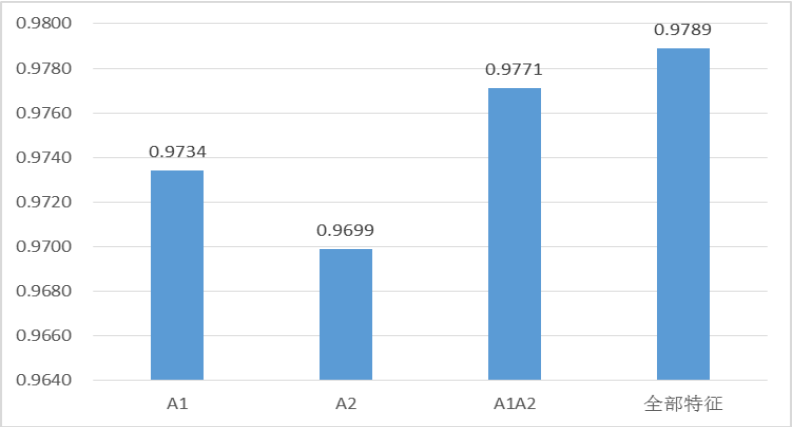

Figure 4. The Influence Feature Set on Prediction Accuracy

(5) Experiment 5:

In this experiment, three other classifiers—MART, KNN and logistic regression are employed to conduct experiments on the same data. Three indicators—precision, recall and F1 score are used to evaluate the effects of the final classification. The results are shown in Table 3-5. It is found out that, in the case of SVM, the precision is 97.89%, the recall rate is 96.54%, and F1 score is 97.03%, which are the optimal values among four algorithms. Thus, SVM algorithm produces the optimum results of prediction.

Table 5. Comparison of Classifier Performance

Methods	Precison	Recall	Macro-F1
MART	0.9144	0.9237	0.9201
KNN	0.9403	0.9358	0.9332
Logistic regression	0.9576	0.9497	0.9526
SVM	0.9789	0.9654	0.9703

4. Conclusion

The main objective of this study is to predict the probability of microblogging forwarding on the basis of analysis on Weibo data. In order to realize this objective, effective feature attributes of Weibo data are extracted, SVM and other machine learning algorithms are adopted for training data and for the generation of prediction model. In the experiments, first, genetic algorithm is applied to assign weights of feature attributes of Weibo, and the larger weighted values of feature attributes are identified as reference of follow-up experiments. Next, in the case of unassigned weights of data, feature attributes are taken out one by one, SVM algorithm is employed to train experiment data and to identify the most influential feature attributes on the final prediction accuracy. After comparing with the previous assigned weights, it is observed that the greater the assigned weight of feature attribute is, the greater the impact on prediction accuracy is after being removed. Then, the correlation among feature attributes is explored. The prediction accuracy is considered not to change significantly after certain two feature attributes are individually removed. Meanwhile, if the removal of the combination of these two feature attributes has a great impact on the experiment results, the correlation between these two feature attributes is proven to be quite close. In the end, four different classifiers are used to unify experimental data. It is found that SVM produces better prediction results in this experiment.

In future study, first, more machine learning algorithms should be introduced, and suitable conditions of different algorithms should be explored. Then, experiments on combinations with three and more feature attributes are to be conducted, and the relation among feature properties is to be further investigated.

References

[1] https://en.wikipedia.org/wiki/Facebook.
[2] https://en.wikipedia.org/wiki/Twitter.
[3] https://en.wikipedia.org/wiki/YouTube.
[4] https://en.wikipedia.org/wiki/Social_network.
[5] Q. Guo, "Journalism and Communication", China Renmin University, **(1999)**.
[6] H. Kwak and C. Lee, "What is Twitter, a social network or a news media? WWW '10 Proceedings of the 19th international conference on World wide web, 2010
[7] D. Boyd, S. Golder and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter", Hawaii International Conference on System Sciences , **(2010)**.
[8] M. S. A. Wolfram, "Modelling the stock market using Twitter", School of Informatics, vol. 74, **(2010)**.
[9] Y. Han, "Research on Key Technologies of Analyzing and Mining Social Networks", National University of Defense Technology, **(2011)**, pp. 1-15.
[10] H. Yao, "Detection and Trend Prediction Research of Hot Topic of Micro-Blogging", South China University of Technology, **(2013)**, pp. 1-8.
[11] J. Yang and S. Counts, "Predicting the speed, scale and range of information diffusion in twitter", The 10th international AAAI conference ON Web and Social Media, **(2010)**.
[12] B. Suh, L. Hong and P. Pirolli, "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network", IEEE International Conference on Social Computing, **(2010)**.
[13] Z. Yang, J. Guo and K. Cai, "Understanding retweeting behaviors in social networks", Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, **(2010)**, pp. 1633-1636.
[14] Y. Li, H. Yu and L. Liu, "Predict algorithm of micro-blog retweet scale based on SVM", Application Research of Computers, vol. 30, no. 9, **(2013)**, pp. 2594-2597.
[15] S. Zhang and W. Cai, "Influence analysis of user characteristics to microblogging retweet behavior", Computer Engineering and Applications, vol. 50, no. 11, **(2014)**, pp. 11-16
[16] Z. Luo, T. Chen and W. Cai, 'Microblogging Retweet Prediction Algorithm Based on Random Forest', Computer Science, vol. 41, no. 4, **(2014)**, pp. 62-74
[17] D. Boyd, S. Golder and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter", Hawaii International Conference on System Sciences, **(2010)**.
[18] Y. Artzi, P. Pantel, and M. Gamon. Predicting responses to microblog posts. The 2012 Conference of the North American Chapter of the Association for Computational, 2012
[19] Z. Luo, M. Osborne, J. Tang and T. Wang, "Who Will Retweet Me?", Finding Retweeters in Twitter. SIGIR'1, **(2013)**.
[20] K. Lee, J. Mahmud and J. Chen, "Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information', Proceedings of the 19th international conference on Intelligent User Interfaces. ACM, **(2014)**, pp. 247-256.
[21] http://www.keenage.com/html/c_bulletin_.htm, **(2007)**.
[22] X. Linhong, L. Hongfei, P. Yu, R. Hui and C. Jianmei, "Constructing the Afective Lexicon Ontology", Journal of the China Society for Scientific and Technical Information, vol. 27, no. 2, **(2008)**, pp. 180-185.
[23] Y. YANG, "An Evaluation of Satisical Apporaches to Text Categorization", Information retrieval, vol. 1, no. 1-2, **(1999)**, pp. 69-9

