ICIC Express Letters
Part B: Applications
Volume **8**, Number **12**, December **2017**

 ISSN 2185-2766
pp. **1643–1651**

INDIVIDUAL FORWARDING PREDICTION ON SOCIAL NETWORKS USING THE KLOUT APPLICATION

Bing Fang and Wenyuan Miao

School of Management
Shanghai University
No. 99, Shangda Road, Baoshan District, Shanghai 200444, P. R. China
melodyfang@shu.edu.cn

Received June 2017; accepted September 2017

Abstract. *The information forwarding function is the fundamental information diffusion mechanism on a social network. This paper aims to predict users' forwarding behavior. Previous studies have usually focused on static information, while ignoring the dynamic interaction and social influence between users. To address this problem, we introduce Klout, an application that measures users' social influence and traces the dynamic interaction between social network users. The research problem is formulated as a classification task, and machine-learning models are developed. In addition to features used to represent users' static information, features that can be used to measure users' social influence are also introduced. These features are constructed from information obtained from the Klout. The experimental data were sourced from Twitter. After data pretreatment, a 10-fold cross validation experiment is adopted. In case of overfitting, a sensitivity test was also conducted under data imbalance situation. The experimental results show that the model with social influence features outperforms the model without such features, which demonstrate the effectiveness and necessity of taking users' social influence into consideration. In addition, the results also demonstrate the effectiveness of the method used to quantify social influence as proposed in this paper.*
Keywords: Individual forwarding, Klout, Social influence, Social network, Machine learning

1. **Introduction.** With the rapid development of online social network sites (SNS) such as Twitter, Facebook and Sina Blog [1], the concept of forwarding prediction has aroused the interest of many researchers. Unlike traditional media, forwarding is the most important mechanism of information diffusion used in social media. Users not only receive information, but also take part in the information diffusion process [2]. At present, most existing studies mainly focus on predicting the forwarding scale of a certain message or topic [3]. This can be called "massive forwarding prediction" and is used for various marketing and public opinion analysis purposes [4,5]. In contrast, very few studies have researched the concept of individual forwarding prediction. Individual forwarding prediction refers to predicting SNS-users' forwarding behavior. Correctly predicting this behavior has various benefits. Firstly accurate predictions can help SNS to provide more personalized services, such as a personalized message recommendation system. Secondly, accurate predictions can help businesses to locate target users among the numerous social network users. Thirdly, accurate predictions of forwarding behavior can help social network users to filter the information [6], as this is an age of information explosion. The reason researchers have not yet investigated this field to any large degree is that individual forwarding prediction aims to identify the specific forwarders and therefore needs to take every single forwarder's characteristics into consideration. This is quite difficult, because there are too many latent forwarders [7]. Researchers involved in this field usually ignore dynamic interaction between users and only focus on users' static information,

Wang et al. [8] predicted users' retweet action based on features of released and accepted users along with content, and then used SVM to filter spam and established a logistic regression to conduct prediction. Tang et al. [9] analyzed retweet features and improved LR for the prediction. Zhou et al. [10] systematically examined the features of followee, follower, tweet, and interaction, and they summarized 52 features in total to perform the prediction. This information on its own is insufficient as a means to predict the degree of a user's social influence. Social psychological studies have demonstrated that social influence could significantly influence user behavior at SNS [11,12].

Given this situation, a novel prediction method of individual forwarding behavior is proposed. Unlike previous studies, Klout is introduced in our proposed method to quantify social influence between users through constructing social-related factors. Klout is a company in which Microsoft invested. Klout aims to use social media analytics to create user profiles and measure the degree of each user's social influence on SNS. They have several technology advantages which may be hard for researchers to realize, given the existing resource constraints. Firstly, Klout created user profiles by aggregating multiple social networks, such as Google+, Instagram, LinkedIn, Twitter, YouTube, and even Wikipedia [14]. Secondly, Klout has the ability to trace each user's information. This includes not only static information, but also each user's activities and interactions. Klout traces about 45 billion interactions on various social networks every day. Thirdly, Klout also has the ability to identify spam/dead accounts and reduce their influence. Klout then labels each user based on the collected data. The labels Klout can apply to any given user are determined as follows.

1) Klout Score: Klout Score is a numerical value, which is used to measure a user's social influence on a given social network. The score ranges from 1 to 100, with 100 being the most influential.
2) Influence and Influences: This information is used to find out which other users can influence a specific user the most, as well as the users who can be influenced most by this specific user. The findings are based on the users' interactions.
3) User Topic: This is used to find the specific topics that interest users. The findings are based on user activities and the technology of natural language processing.

In sum, the main technical achievements of our work are threefold.

1) We design a novel method to predict individual forwarding behavior.
2) We introduce Klout in the field of forwarding prediction through quantifying social influence between users. Our findings indicate that it is feasible to predict individual forwarding behavior with the consideration of features constructed from Klout.
3) Our work reveals the law of information diffusion between individuals on SNS, which builds a solid foundation for further study of the whole process of information diffusion on SNS.

This distinguishes our work from existing studies that fail to systematically investigate structure properties of SNS. In this paper, we adopt Klout technique and machine learning methods to quantify social influence between users and predict the individual forwarding behavior respectively. To test the performance of our proposed method, we form the baseline method by concluding basic features from previous literature, and compare it with our proposed method.

The rest of this paper is organized as follows. Section 2 introduces the model construction. Section 3 describes the dataset used in the experiment. The results and discussion are introduced in Section 4. Section 5 presents our conclusions.

2. **Model Construction.**

Four categories of features (profile features, content features, structural features, and Klout features) are used in the proposed model.

Profile features: This group of features is used to measure the activity of receivers and publishers. This category includes the number of followers, number of people being followed, number of published messages and the frequency of publication. If a user has more followers (number of followers), follows more users (number of people being followed), publishes more messages (number of messages), and publishes frequently (frequency of publication), the user will be more active in social media. Such users are more likely to be engaged in the process of information diffusion. All the above features can be acquired from a user's profile page.

Content features: This group of features reflects the characteristics of a user's message content, including whether the message contains the character "#" (which means the topic); whether the message contains the character "@" (which means the message is directed @ someone); whether the message contains a URL (which means the message has multimedia characteristics), and the length of the message.

Structural features: These features reflect the structural characteristics of social media users. Previous studies indicate that a user's structural characteristics relate closely to that user's behavior on a social network [12]. Here, the local cluster coefficient, page rank, and the rate of common friends were chosen.

The local cluster coefficient reflects the density of connections between each information receiver's friends. Online social networks clearly exhibit clustering effects. Previous studies have shown that the density of connections of a specific user's friends makes that user more involved in the social network [12]. The formula is as follows:

$$\text{ClusterScore} = \frac{\text{Number of connections among A's friends}}{N(N-1)/2} \tag{1}$$

Here, N is the total number of friends of user A.

PageRank is a well-known structural feature, which initially was used to rank websites. Now many researchers employ PageRank to represent a user's degree of social influence on the global network [7].

In addition, the rate of common friends was also constructed, in order to evaluate the social influence between information receivers and information publishers. A user's behavior is often influenced by his or her friends [13]. If user B has more friends who follow user A, then user A may have more influence than user B. Therefore, the common friend rate is constructed, and the formula is as follows:

$$\text{CFR} = \frac{\text{Number of common friends}}{\text{Number of information receivers' friends}} \tag{2}$$

All of the above features can be acquired through a user's static connection information.

Klout-based features: This group of features is constructed from Klout, in order to measure the degree of social influence between receivers and publishers. These features include Klout score, topic similarity, mutual influence index, influence similarity and influencer similarity. Unlike structural features, which mainly focus on static connection information, Klout-based features are constructed from a user's dynamic interactions.

A Klout score is a numerical value, which is used to measure a user's degree of social influence on a social network. The score ranges from 1 to 100, with 100 being the most influential. The calculation of the score is based on features aggregated from multiple dimensions of a user's social network interactions. The score shows the user's degree of global social influence on a social network. Users with higher scores are able to spread information more effectively within the network [14].

Topic similarity: Previous studies have demonstrated that information receivers prefer to forward messages from users who share information on similar topics [7]. Klout has the ability to exhibit each user's topic preferences. Cosine similarity is applied, in order

to calculate the topic similarity between two users. The formula is as follows:

$$\text{TopicSim(A, B)} = \frac{\text{topic\_A} \cdot \text{topic\_B}}{\|\text{topic\_A}\|\|\text{topic\_B}\|} \tag{3}$$

Here, topic_A, topic_B represent the topic arrays of user A and user B, respectively.

"Influencee similarity" and "influencer similarity" reflect the closeness of the social circle of information receivers and publishers on the social network. The degree of closeness is also calculated by cosine similarity.

The mutual influence index is produced by the following rules:

Suppose user A is the information receiver, and B is the publisher.

Initial: Mutual influence index = 0

(1) If (A is in B's influencer list): Mutual influence index +=1.
(2) If (B is in A's influencee list): Mutual influence index +=1.
(3) If (B is in A's influencee list, and A is in B's influencer list): Mutual influence index +=1.

3. **Dataset Description.**

3.1. **Data collection.** The data were collected from Twitter, one of the most popular social media sites in the world. Forwarded messages on Twitter can also be referred to as retweets. Initially, 321 seed users were acquired by randomly choosing one user on Twitter and then obtaining that user's friends list, which contained 320 users. From these seed users, a whole network was trawled. That network contains 17,682 users, including the initial 321 seed users and an additional 17,361 users who are in the range of one hop of the seed users. Furthermore, tweets posted by trawled users were collected from March 21, 2016, to July 7, 2016.

3.2. **Data balance.** For our forwarding prediction evaluation data set, each project represents a forwarding action. The action is marked with the class label "1" if a certain tweet is retweeted by a certain user. Otherwise, the action is marked with the class label "0". However, a severe data imbalance exists, because the number of non-forwarding actions is much higher than the number of forwarding actions, at a ratio of 1:12. This finding is consistent with the ratio in Morchid M's work [15]. Traditional methods of data balancing are over-sampling and under-sampling. However, both methods may fail in a forwarding prediction scenario, because both methods assume that all the tweets published by a user's followers will be browsed, but actually it is impossible for users to browse all tweets on their home pages. To address this problem, we employed a technical means to select non-retweeted tweets. The main steps of this method are as follows.

Step 1: Retrieve the friend list of a user from the database.

Step 2: Retrieve his or her friends' published tweets from the database.

Step 3: Sort the tweets chronologically.

Step 4: Locate and mark the retweeted tweets.

Step 5: Locate and mark the non-retweeted tweets, which are the four nearest tweets to each retweeted tweet.

Repeat Step 1 to Step 5 for every user.

The first three steps reconstruct the home page of each user. The remaining two steps choose data. The nearest non-retweeted tweets to each retweeted tweet are chosen, because compared with other non-retweeted tweets, the probability of these tweets being browsed by the user is undoubtedly much higher.

After data balancing, the amount of retweeted tweets and chosen non-retweet tweets were 56,182 and 223,728, respectively.

4. **Results and Discussion.** An experimental study was designed to test the proposed method by comparing our method with alternative methods, which did not consider Klout-based features in previous studies. All features used in both methods are summarized in Table 1. The difference is that Klout features are only used in the proposed method.

Table 1. Summary of features

Category	Features
Profile Features	Number of followees
	Number of followers
	Number of collection
	Number of tweets
	Frequency of publication
Content Features	IsTopic
	IsAlt
	Length of tweet
	IsUrl
Structural Features	Local cluster coefficient
	Common friend rate
	Page rank
Klout Features	Klout
	Topic similarity
	Mutual influence index
	Influencee_sim
	Influencer_sim

Several conventional classifiers were chosen to perform the classification, including support vector machine (SVM) and logistic regression (LR). Besides, ensemble classifiers such as random forest (RF) and Adaboost (ADA) were also included for comparison purposes.

In order to obtain robust results, a 10-fold cross validation experiment was adopted. Here, the amount of retweeted tweets is equal to non-retweeted tweets (Figure 1).

The results demonstrate that the proposed method significantly outperforms the alternative method, with the average increment approximately 11.4% for F1-Measure. The best-performing classifier is random forest, whose F1-Measure is from 66.1% to 79.8%, with the increment of up to 20.7%.

To test the sensitivity of both methods to the rate of retweeted tweets and non-retweeted tweets, a hold-out procedure was conducted, with the ratio of retweeted tweets and non-retweeted tweets changing from 1:1 to 3:1, and the training set and testing set fixed to 4:1. The results of these three experiments are shown in Figure 2.

As shown in Figure 2(a), considering the precision, the proposed method performs slightly better than the alternative method. Meanwhile, it can be seen that when the ratio is 3:1, the precision achieves the highest. That is because the more non-retweeted tweets are used, the more classifiers tend to classify tweets into "non-retweeted". The precision can achieve 75% if classifiers classify all tweets into "non-retweeted" under the ratio of 3:1. Hence, as a supplement, recall is used and the result is shown in Figure 2(b), and the proposed method performs much better than the alternative method at any ratio. Meanwhile, in contrast to precision, it can also be seen that when the ratio is 1:1, the recall achieves the highest. When considering the F1-Measure, as shown in Figure 2(c), the proposed method performs better than the alternative method at any

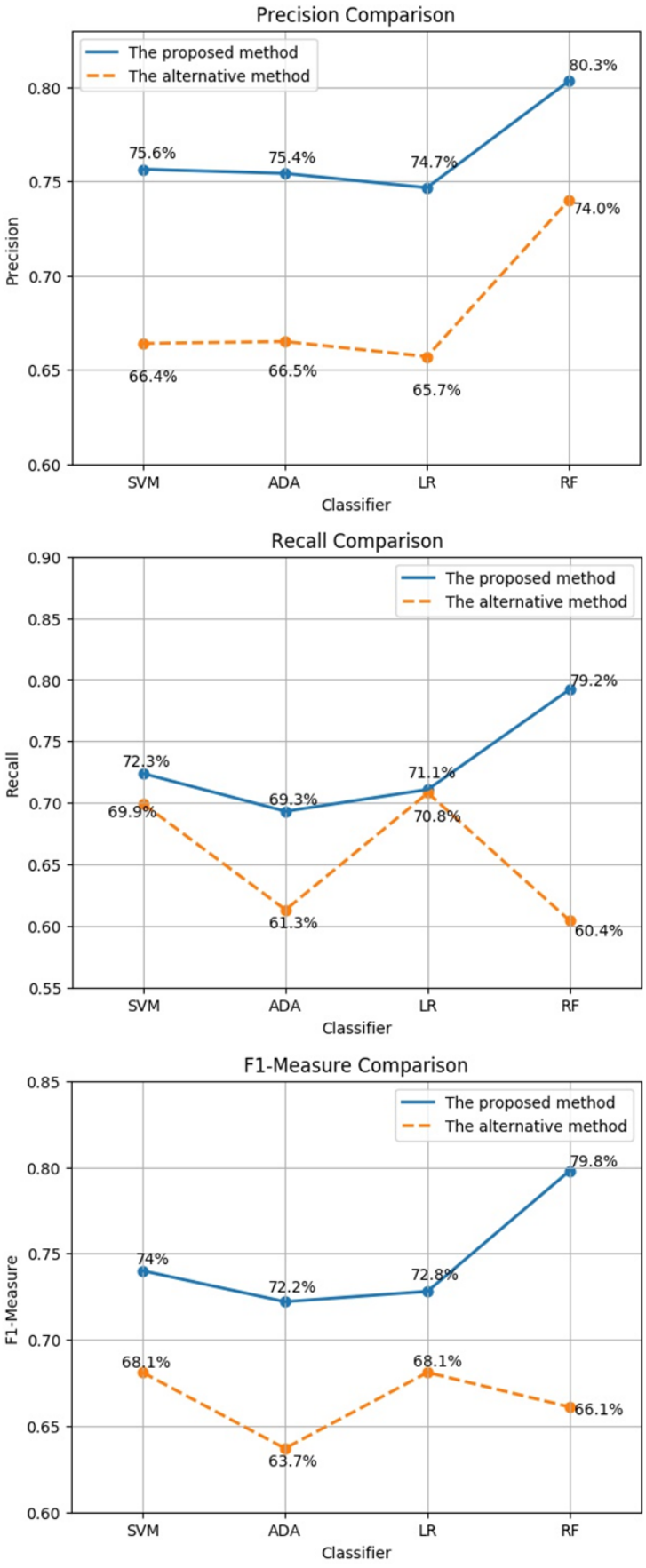
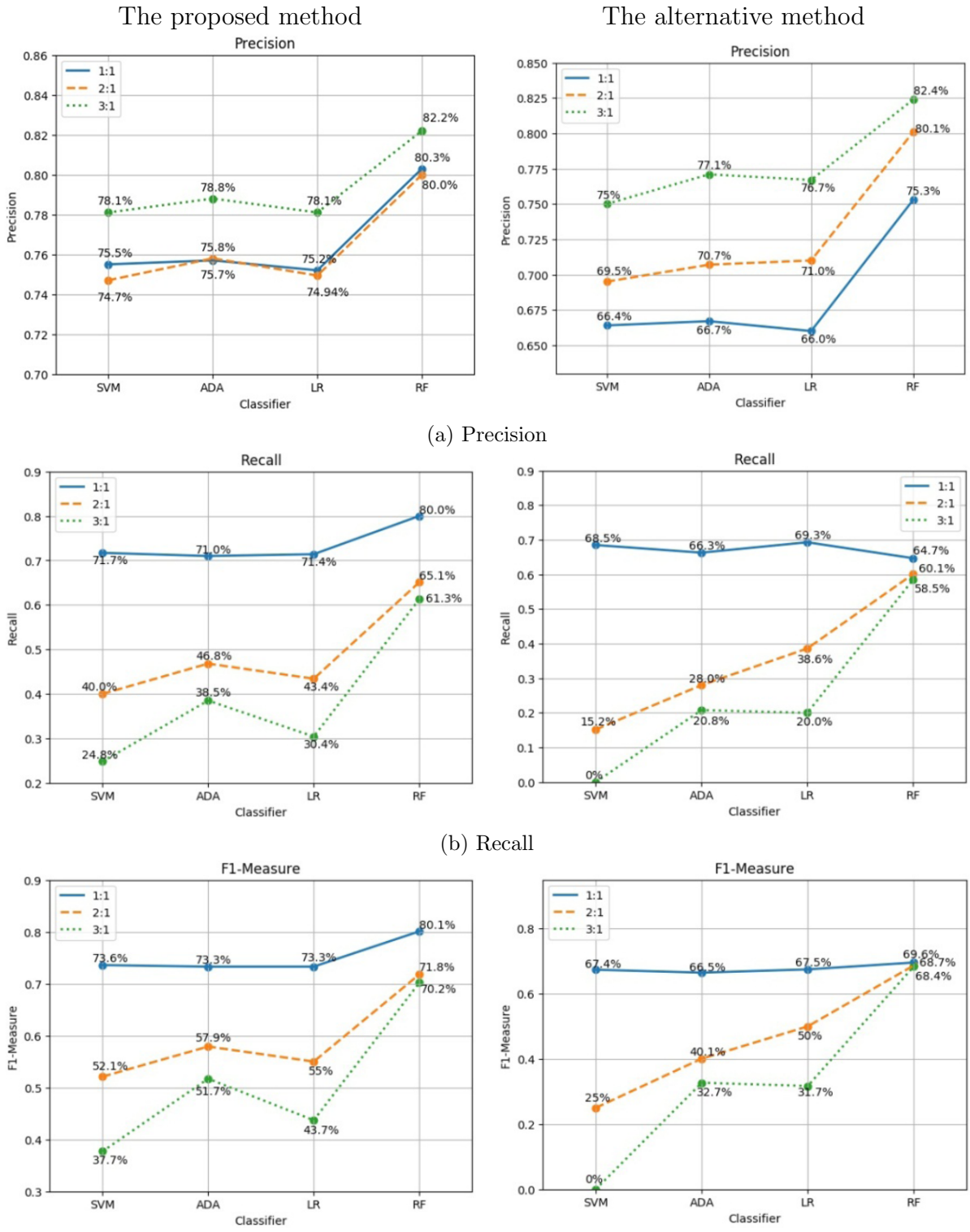


FIGURE 1. Results of 10-fold cross-validation



(a) Precision

(b) Recall

(c) F1-Measure

FIGURE 2. Results of the sensitivity test

ratio. Combining the results shown in Figure 2, we can come to the conclusion that our propsosed method outperforms the alternative method.

Meanwhile, one point of interest is that the F1-Measure of SVM in the alternative methods comes to zero at the ratio of 3:1. This happens because the main idea of SVM is to "maximize the distance between the hyperplane and the points that are closer to it, and SVM has a bias toward the majority class" [16]. At the ratio of 3:1, SVM simply

classifies all projects into "Class 0" which causes the recall and F1-Measure to be zero in the alternative methods. Meanwhile, in the proposed method, the F1-Mesausre of SVM is still 37.7%, which also demonstrates the efficiency of our proposed methods.

In this paper, the degree of a user's social influence is introduced by means of constructing social influence features based on information obtained from Klout. The results show that introducing social influence into individual forwarding predictions is both effective and necessary. However, some limitations still exist in the proposed model, because the Klout application only supports users of Twitter, Google+ and Instagram. For this reason, the degree of social influence of users on other social network sites cannot be obtained through our proposed method.

5. **Conclusions.** In this paper, a machine learning model is proposed for purposes of individual forwarding prediction. Our model takes the degree of a user's social influence into consideration. Features constructed out of information obtained from Klout are added to the model. Both 10-fold cross-validation experiments and the sensitivity test demonstrate the superiority of the proposed model. This paper sheds light on the importance of researchers who study user behavior on SNS taking social influence into consideration. Our study also provides an approach that can be used to quantify the degree of social influence with the Klout application.

In the future, we will carry out more experiments on social network sites other than Twitter, and we will dynamically analyze users' information obtained from Klout to adapt changes in time.

Acknowledgment. This work is partially supported by the Shanghai Natural Science Foundation of China (Project No. 16ZR1447100) and Shanghai Junior Faculty Cultivation Plan at Universities (Project No. N.37-0129-15-201).

REFERENCES

[1] L. Hong, O. Dan and B. D. Davison, Predicting popular messages in Twitter, *International Conference on World Wide Web*, Hyderabad, India, pp.57-58, 2011.

[2] W. Hou, Y. Huang and K. Zhang, Research of micro-blog diffusion effect based on analysis of retweet behavior, *International Conference on Cognitive Informatics & Cognitive Computing*, pp.255-261, 2015.

[3] E. Kim, Y. Sung and H. Kang, Brand followers' retweeting behavior on Twitter: How brand relationships influence brand electronic word-of-mouth, *Computers in Human Behavior*, vol.37, pp.18-25, 2014.

[4] K. H. Chu et al., Diffusion of messages from an electronic cigarette brand to potential users through Twitter, *Plos One*, vol.10, no.12, 2015.

[5] X. Lu et al., Predicting the content dissemination trends by repost behavior modeling in mobile social networks, *Journal of Network & Computer Applications*, vol.42, no.3, pp.197-207, 2014.

[6] W. M. Webberley, S. M. Allen and R. M. Whitaker, Retweeting beyond expectation: Inferring interestingness in Twitter, *Computer Communications*, vol.73, pp.229-235, 2016.

[7] X. Tang, Q. Miao, Y. Quan, J. Tang and K. Deng, Predicting individual retweet behavior by user similarity, *Knowledge-Based Systems*, vol.89, pp.681-688, 2015.

[8] Z. Wang, K. Liu and Z. Zheng, Prediction retweeting of microblog based on logistic regression model, *Journal of Chinese Computer Systems*, vol.37, no.8, pp.1651-1655, 2016.

[9] X. Tang, Y. Quan, J. Song et al., Novel algorithm for predicting personalized retweet behavior, *Journal of Xidian University*, 2016.

[10] J. Zhou et al., Predicting who will retweet or not in microblogs network, *Chinese National Conference on Social Media Processing*, Springer, Singapore, 2015.

[11] S. P. Borgatti et al., Network analysis in the social sciences, *Science*, vol.323, no.5916, pp.892-895, 2009.

[12] Z. Katona, P. P. Zubcsek and M. Sarvary, Network effects and personal influences: The diffusion of an online social network, *Journal of Marketing Research*, vol.48, no.3, pp.425-443, 2011.

[13] J. Zhang et al., Who influenced you? Predicting retweet via social influence locality, *ACM Trans. Knowledge Discovery from Data (TKDD)*, vol.9, no.3, p.25, 2015.

[14] A. Rao et al., Klout score: Measuring influence across multiple social networks, *IEEE International Conference on Big Data*, pp.2282-2289, 2015.

[15] F. Song, Z. Guo and D. Mei, Feature selection using principal component analysis, *International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM)*, vol.1, 2010.

[16] L. Smallman and A. Artemiou, A study on imbalance support vector machine algorithms for sufficient dimension reduction, *Communication in Statistics – Theory and Methods*, vol.6, 2017.