ICIC Express Letters
Part B: Applications
Volume **10**, Number **5**, May **2019**

ICIC International ©2019 ISSN 2185-2766
pp. **371–378**

PREDICTING RETWEET BEHAVIOR ON SOCIAL MEDIA WITHIN COMMUNITIES FROM THE PERSPECTIVE OF USER BEHAVIOR SPREADING

Liuqing Meng[1], Bing Fang[1,*], Xiaoying Liu[1], Yibo Shang[1] and Ang Luo[2]

[1]School of Management
Shanghai University
No. 99, Shangda Road, Shanghai 200444, P. R. China
{ mengliuqing; vickyliu0923; shangyibo }@shu.edu.cn; *Corresponding author: melodyfang@shu.edu.cn

[2]SINA Plaza
No. 8, Xibeiwang East Road, Haidian District, Beijing 100193, P. R. China
luoang@staff.sina.com.cn

Received October 2018; accepted December 2018

Abstract. *Retweet prediction on social media is an important task for studying the principle of information diffusion. However, most researches focus on retweet prediction at the individual or site-wide levels and they ignore community structure. Previous studies have examined various factors influencing user retweet behavior, such as content influence and user influence. Yet a unified framework has not been formed. The main contribution of our research work is the design of a novel framework for predicting the retweet behavior of social media users from the perspective of user behavior spreading. We consider the three influences on user retweet behavior: spontaneous behavior, information influence, and user susceptibility. Based on real-world social media data, our experimental results reveal that the proposed framework outperforms existing methods in terms of three evaluation metrics. This proposed methodology can be applied to accurately pinpointing audiences and improving the efficiency of advertising.*
Keywords: Retweet predication, User behavior spreading, Social media, Information diffusion

1. **Introduction.** On social media, information is disseminated to participants through user retweet behavior. Understanding the mechanisms of social media information dissemination and predicting retweet behavior are essential for various applications, such as user behavior analysis, business intelligence, and popular event prediction [1].

At present, studies predicting retweet behavior mainly focus on two aspects: massive retweet prediction [2,3], and individual retweet prediction [4,5]. However, most microblogs are only popular among small groups among users. Social networks are made up of many communities, and people with similar interests, backgrounds, attitudes, and values spontaneously form communities [6]. According to social psychology researches, the actions of people in the same community are homogeneous, meaning that people tend to follow the social networking actions of their friends and others in their communities. In fact, information dissemination on social networks passes through three stages. First, a post is published in one community, and then it spreads quickly across other communities, and finally becomes a hot topic at the level of the whole social networking site (SNS). At the second stage, most microblog posts stop spreading. However, most research ignores the second stage, thereby glossing over the community structure of the SNS. To deepen our understanding of information diffusion on social networks, our study concerns the second stage: information popularity in different communities.

DOI: 10.24507/icicelb.10.05.371

On social media, the spread of information is just a process of user behavior spreading: a user observes someone reposting a microblog post, and he or she also retweets the post. The spread of user behavior on social networks has attracted attention from several scholars [7-9]. Aral found that user behavior spreading was affected by three factors: spontaneous behavior, influence, and susceptibility. More importantly, all three factors must be taken together to predict the propagation of behaviors.

Following Aral's work, we divide the factors affecting user retweet behavior into three categories: user spontaneity, information influence, and user susceptibility. Instead of studying these factors separately, as previous research has done, we discuss the interactions among the factors. In this work, we convert the prediction of user retweet behavior within specific communities into a classification task and select appropriate variables from the three factors. All the variables are input into different classifiers to test the accuracy of our model.

The rest of the paper is organized as follows. In Section 2, we describe in detail the proposed theoretical framework and selection of features. The experiment is presented in Section 3. Results and discussion are described in Section 4. Finally, we conclude this paper in Section 5.

2. Research Framework.

Our research can be divided into three parts: community detection, feature selection, and classification.

2.1. Community detection.

The aim of our research is to predict whether the information will become popular within communities. Therefore, dividing users into different communities is the first task. People belonging to the same community are tightly connected, whereas people from different groups are sparsely connected [10]. A popular modularity approach is the Louvain method, which iteratively optimizes local communities until global modularity can no longer be improved given perturbations to the current community state.

2.2. Feature selection.

This study analyzes influencing features from the perspective of user behavior and classifies the features into three categories: 1) spontaneous behavior, 2) influence of information on community members and 3) the influence among members of the community.

2.2.1. *Spontaneous behavior.* It refers to the extent to which people like reposting and commenting. It is represented by community activity and the formula is as follows:

$$community\ activity = \frac{\sum_0^n zf\_num + pl\_num}{n} \tag{1}$$

where zf_num is one individual's total number of reposts, pl_num is one individual's total number of comments, and n is the total number of members in the community.

2.2.2. *Influence of information on community members.* Combining forces of information influence and user susceptibility, we consider this factor from three aspects.

1) *Influence of content on community members.* We measure this factor by the semantic similarity: Google Word2Vec model [11] is applied to measuring semantic similarity between two microblogs.

2) *The influence of content creator on community members.* It can be reflected by the network topology structure between the content creators and the community members. Most prior research has focused on the characteristics of content creators and receivers separately, ignoring the connections between them. In this study, we measure the connections by analyzing (a) the social influence of the content creator on the whole network and (b) the relationship between the content creators and the receivers.

(a) *The social influence of content creator on the whole network.* PageRank is a well-known structural feature that was used to represent users' social impact on networks [4,12]. The formula is as follows:

$$PageRank(u_i) = d + (1+d) \sum_{V_j \in I(u_i)} \frac{PageRank(U_j)}{O(V_j)} \tag{2}$$

where $PageRank(u_i)$ is the influence of user i, $I(u_i)$ is the follower set of u_i, $O(V_j)$ is the number of V_j's followers. The damping d is often set as 0.15 to make the final result converge.

(b) *The relationship between the content creator and the receivers.* Whether the content creator and the receiver are in the same community is also an influencing factor, because the actions of people in the same community are homogeneous.

3) *The joint influence of content and content creator.* Previous studies have discussed the influence of content and that of content creator separately. However, that is not enough. Owing to different degrees of user trust in each microblog post, people will only retweet certain microblog post released by a content creator, instead of all posts. Therefore, we need to consider the interaction between content and content creator, examining these factors simultaneously. Like with semantic similarity, we measure this joint influence by summing the similarity between the new microblog post released by a content creator and previous microblog posts published by the same content creator and retweeted in the same community.

2.2.3. *The influence among members of the community.* In addition to the external influence on a community, members of a community will also produce mutual forces [13]. On social networks, the connections among people are modeled using network diagrams. Centrality is used to illustrate the importance of nodes in the network, and centralization to illustrate the influence of the whole network. The overall influence of the community can be reflected by the network structure and the tightness of connections among members.

Previous research has considered the influence of the content creator from the whole-network perspective. In our study, we measure the influence of the content creator at the community level, determining the tightness of connections between the content creator and the community members with the following six features.

Degree Centrality. It refers to the number of links connecting to a node. It reflects the user's influence in the network and information dissemination capabilities. Its formula is:

$$C_D(N_i) = \frac{\sum_{j=1}^{n} X_{ij}(i \neq j)}{n-1} \tag{3}$$

where $C_D(N_i)$ represents the degree of node i, n is the total number of users in the community, $X_{ij} = 1$ if there is a connection between user i and user j, else $X_{ij} = 0$.

Betweenness Centrality. It is the number of shortest paths that pass through a node. For the entire network, users with large betweenness may be connected to two or more community hub nodes, playing a key role in the dissemination of information in the entire network. Its formula is as follows:

$$C_B(N_i) = \frac{\sum_{i}^{n} \sum_{j}^{n} b_{ij}(N_i)}{n(n-1)/2} \tag{4}$$

where $\sum_{j}^{n} b_{ij}(N_i)$ means the length of the shortest path of connection between node i and node j.

Closeness Centrality. Closeness centrality of a node is the average length of the shortest path between the node and all other nodes in the graph. The closer users are to others, the more they do not rely on others in the process of disseminating information. Its

formula is as follows:

$$C_C(N_i) = \left[\frac{\sum_{i=1}^{n} d(N_i, N_j)}{n-1}\right]^{-1} \tag{5}$$

where $d(N_i, N_j)$ is the length of the shortest path between node i and node j.

Eigenvector Centrality. Similar to PageRank, eigenvector centrality is a measure of the influence of a node in a network. The eigenvector centrality is calculated from the sum of the eigenvector centralities of adjacent nodes [14], and its formula is as follows:

$$C_E(N_i) = \sum_{i=1}^{n} a_{ij} e(N_j) \tag{6}$$

where $a_{ij} = 1$ if node i is linked to node j, and $a_{ij} = 0$ otherwise.

Cluster Score. It is the degree to which nodes tend to cluster together. A higher cluster score means there exist denser social links among a user's friends, and their information is thus more likely to be retweeted by their friends. Its formula is:

$$ClusterScore = \frac{C\_F}{N(N-1)/2} \tag{7}$$

where N is the number of total friends of user i and C_F is the number of connections among user i's friends.

The Overall Influence of the Community. According to the topological structure analysis of social networks, the topology structure of a community can be represented by four indicators: degree centralization [15], betweenness centralization [16], closeness centralization [17], and clustering coefficient [18], which show the aggregation degree for degree centrality, betweenness centrality, closeness centrality, and cluster score of the community, respectively. Its formula is as follows:

$$C = \frac{\sum_{i=1}^{n} c_{\max} - c_i}{\max\left[\sum_{i=1}^{n} (c_{\max} - c_i)\right]} \tag{8}$$

where $c_{\max}$ refers to the maximum centrality of the network, and n is the total number of users of the community.

2.3. **Classification.** In this paper, we aim to predict microblog posts' popularity within communities. We formulate the prediction problem as a binary classification task. In previous studies, most researchers have identified popular microblogs according to retweet count [3,19-21]. We likewise divide microblogs into popular microblogs and unpopular microblogs based on the number of times posts on these microblogs have been retweeted by the community. We choose five supervised machine learning methods to perform the classification: 1) support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection; 2) random forests are an ensemble learning method for classification that operate by constructing a multitude of decision trees; 3) gradient boosting produces a prediction model in the form of an ensemble of weak prediction models and generalizes them by allowing optimization of an arbitrary differentiable loss function; 4) AdaBoost can be used in conjunction with many other types of learning algorithms to improve performance; 5) multi-layer perceptron algorithm is a neural network algorithm which imitates the structure and function of biological neural network and is used to estimate or approximate functions.

3. **Experiment.** In order to verify the superiority of our model, we compare our model with previous methods. Previous studies mostly focused on the prediction of the popularity in the whole network, while our study is to predict the popularity within the community and choose features different from previous studies. The details will be described in 3.3. The whole experimental process is divided into the following four parts.

3.1. **Data collection.** Our dataset was collected from Sina-Weibo. The dataset consists of 648,830 users and 10,237,045 social links between them.

3.2. **Community detection.** After obtaining users' data, we applied the Louvain method to detecting communities from the social graph. Figure 1 and Figure 2 show examples of a small community and a big community. The vertices represent users and the lines represent connections. We removed communities with less than 1,000 or more than 20,000 members because large-scale or small-scale communities are not typical in the real world. Thus, the final dataset consisted of 40 communities with 234,180 users. These users published 116,535 original microblog posts. Further, information related to the microblog posts, such as the content, the content creator, the number of reposts, the number of likes, and the number of comments was collected. The number of members in each community is shown in Figure 3.

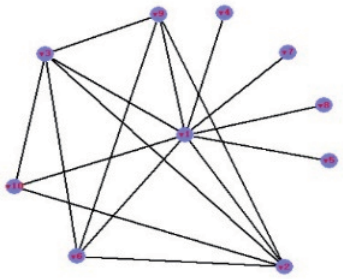


FIGURE 1. A small-scale community

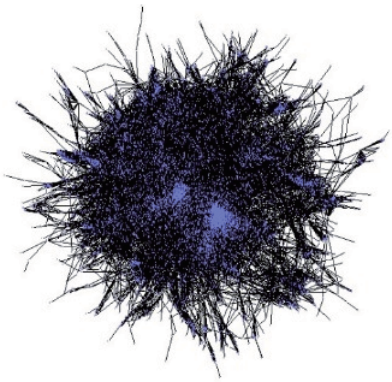


FIGURE 2. A large-scale community

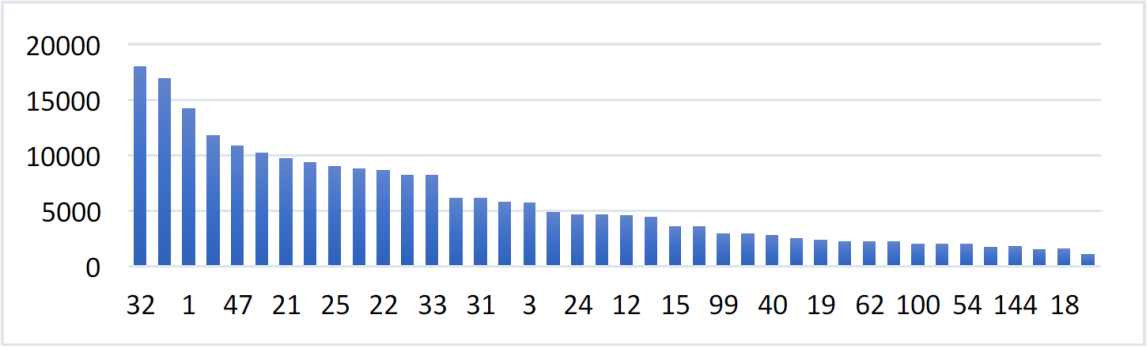


FIGURE 3. The number of members of communities

3.3. **Feature construction.** According to previous research, there are four types of factors affecting user retweet behavior: 1) the form of information (whether a microblog post contains URL, "#", "@" and its length); 2) the sentiment of information (the sentiment score is 1 for positive content, −1 for negative, and 0 for neutral); 3) the content of information (semantic similarity); 4) the influence of information content creator (calculate individual influence through PageRank). To examine the validity of our experimental group, we used those four features, which consist of seven variables, to define our baseline group (Table 1).

According to our novel proposed framework, we identify twelve new features to model sharing behavior at the community level. These twelve features, the variables for our experimental group, are shown and described in Table 2.

TABLE 1. Features of the baseline group

Features	Descriptions
isUrl, isAlt, IsTopic, length	The form of information
SentimentScore	The sentiment of information
Sim1	The content of information
PageRank	The influence of information content creator

TABLE 2. Features of the experimental group

Features	Descriptions
Community activity	Spontaneous behavior
IsIncommunity	The relationship between the content creator and receivers
Sim2	The joint influence of content and content creator
Degree Centrality, Betweenness Centrality, Closeness Centrality, ClusterScore, Eigenvector Centrality	The connections between the content creator and community members
Degree Centralization, Closeness Centralization, Betweenness Centralization, Clustering Coefficient	The overall influence of the community

3.4. **Classification.** Among all the microblog posts retweeted in a given community, we define microblog posts whose number of retweets ranks in the top 10% as popular, and mark them as 1, otherwise as 0. To balance the data, we randomly chose, from among all the unpopular posts, a number of unpopular posts equal to the number of popular posts. Thus, the final dataset consists of 34,000 records, with 17,000 popular microblog posts and 17,000 unpopular posts. We use Scikit-learn library for the Python programming language, including support vector machines (SVM), random forests (RF), gradient boosting (GDB), ada boosting (ADA), and multi-layer perceptron (MLP). In order to get robust results, five-fold cross-validation was adopted.

4. **Results and Discussion.** To evaluate the effectiveness of our proposed framework, we apply three common performance measures: precision, recall, and F1-Measure. It is illustrated by the confusion matrix, which consists of false positive (FP), false negative (FN), true positive (TP), and true negative (TN). Based on the confusion matrix, three common performance measures were defined as follows.

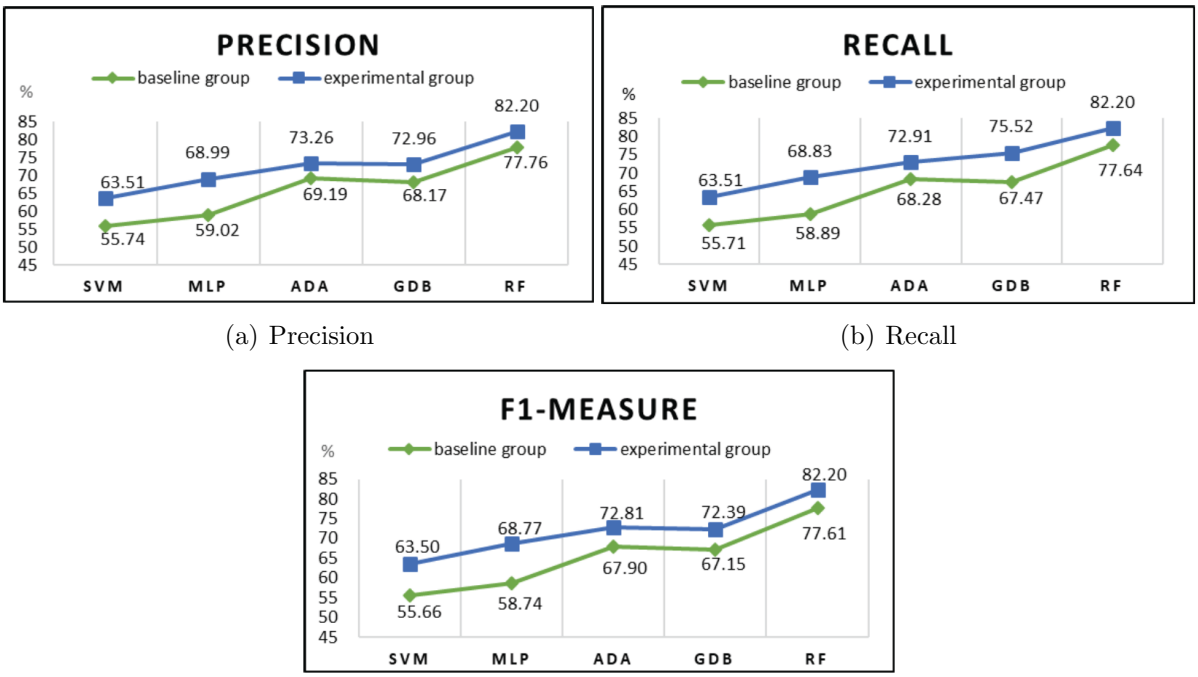
$$Precision = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{10}$$

$$F1\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\% \tag{11}$$

From Figure 4, we can see that our proposed method performed better than the baseline group. Our model's precision figure was greater than that of the baseline group by about 6.23%. Of the machine learning methods tested, MLP yielded notable improvement, from 59.02% to 68.99%. RF had the best precision result, with 82.20%.

With respect to recall, our proposed method gave a greater increase compared with the augment in precision; the average increase was about 6.99%. Further, the most prominent



(a) Precision

(b) Recall

(c) F1-Measure

FIGURE 4. Three performance measures

improvement was from 58.89% to 68.83%, also coming from MLP, and the best recall was 82.20%, also obtained by RF.

With respect to F1-Measure, we find that our proposed method always outperformed the baseline method. The average increase was about 6.16%. The greatest improvement was from 58.74% to 68.77%, coming from MLP, followed by SVM and GDB, with 7.84% and 5.24%, respectively. The best F1-Measure was 82.20%, obtained by RF.

In summary, compared with the baseline features, our proposed method with new features performs better in all three evaluation metrics. As for classifiers, we find that the best classification results originated from the RF method for all evaluation metrics. Its F1-Measure was 82.20%. SVM performed worst for all three metrics. MLP and GDB showed the most improvement when using our proposed method. Considering that the F1-Measure comprehensively reflects recall and precision, the ensemble methods, RF and GDB, performed better than single classifiers.

5. **Conclusion.** In this paper, we study the retweet behavior of users on social media within communities. Our main contribution is to propose a novel framework from the perspective of user behavior spread. Specifically, we divide the factors that affect microblog retweet behavior into three aspects: spontaneous behavior, information influence, and user susceptibility. First, we apply the technique of community detection based on social network analysis. Then, we construct different features based on our proposed framework. Natural language processing technology, complex network analysis technology, and the Word2Vec model are used for feature calculation. Baseline features are chosen following previous research. Finally, we trained five classification methods and evaluated the performance of these algorithms. The result shows that predictions using our proposed features outperform those of previous research.

However, we only studied the Sina-Weibo, and it is necessary to analyze data from other platforms for evaluating the performance of our model. Additionally, we used manual annotation to calculate sentimental score. In future research, we will discuss the

sentimental polarity of Chinese text further. Finally, the dynamic propagation process of microblogging will be studied, in which retweet time will be considered.

Acknowledgement. This work is partially supported by Shanghai Natural Science Foundation of China (No. 16ZR1447100) and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 18YJC860006).

REFERENCES

[1] J. Chen, Y. Liu and M. Zou, User emotion for modeling retweeting behaviors, *Neural Networks*, vol.96, pp.11-21, 2017.
[2] B. Jiang et al., Retweet prediction using social-aware probabilistic matrix factorization, *International Conference on Computational Science*, pp.316-327, 2018.
[3] M. Morchid et al., Feature selection using principal component analysis for massive retweet detection, *Pattern Recognition Letters*, vol.49, pp.33-39, 2014.
[4] X. Tang et al., Predicting individual retweet behavior by user similarity, *Knowledge-Based Systems*, vol.89, pp.681-688, 2015.
[5] Y. Xiao et al., Who will retweet? A prediction method for social hotspots based on dynamic tensor decomposition, *Science China Information Sciences*, vol.61, pp.98-105, 2018.
[6] V. D. Blondel et al., Fast unfolding of communities in large networks, *Journal of Statistical Mechanics*, vol.2008, no.10, pp.155-168, 2018.
[7] S. Aral and D. Walker, Identifying influential and susceptible members of social networks, *Science*, vol.337, no.6092, p.337, 2012.
[8] M. Liu and L. Wang, Analysis and prediction of microblog user behavior on social networks, *Journal of Taiyuan University of Technology*, 2016.
[9] X. Li and Y. Zhao, Analysis of social network user behavior, *Computer Era*, 2017.
[10] V. D. Blondel, J.-L. Guillaume and R. Lambiotte, Fast unfolding of communities in large networks, *Journal of Statistical Mechanics Theory & Experiment*, 2015.
[11] T. Mikolov et al., Distributed representations of words and phrases and their compositionality, *International Conference on Neural Information Processing Systems Curran Associates Inc.*, vol.26, pp.3111-3119, 2013.
[12] Y. Zhang, H. Zhang and W. Zhang, Quick ranking algorithm for network user based on power law distribution, *Journal of Chinese Information Processing*, vol.26, no.4, pp.122-128, 2012.
[13] A. J. Morales et al., Efficiency of human activity on information spreading on Twitter, *Social Networks*, vol.39, no.1, pp.1-11, 2014.
[14] Y. Wang et al., A complex network-based importance measure for mechatronics systems, *Physica a-Statistical Mechanics and Its Applications*, vol.466, pp.180-198, 2017.
[15] J. M. Mccullough, E. Eisencohen and S. B. Salas, Partnership capacity for community health improvement plan implementation: Findings from a social network analysis, *BMC Public Health*, vol.16, no.1, p.566, 2016.
[16] M. W. Schoen et al., Social network analysis of public health programs to measure partnership, *Social Science & Medicine*, vol.123, pp.90-95, 2014.
[17] M. J. Kim, H. Ahn and M. J. Park, A theoretical framework for closeness centralization measurements in a workflow-supported organization, *KSII Trans. Internet & Information Systems*, vol.9, no.9, pp.3611-3634, 2015.
[18] S. J. Hardiman and L. Katzir, Estimating clustering coefficients and size of social networks via random walk, *International Conference on World Wide Web*, pp.539-550, 2013.
[19] Q. Deng et al., Prediction of retweet counts by a back propagation neural network, *Journal of Tsinghua University*, 2015.
[20] 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.
[21] W. M. Webberley, S. M. Allen and R. M. Whitaker, Retweeting beyond expectation: Inferring interestingness in Twitter, *Computer Communications*, vol.73, no.6, pp.229-235, 2015.