Journal of Computers Vol. 29 No. 4, 2018, pp. 147-160
doi:10.3966/199115992018082904012

Retweet Prediction within Communities on SNS Based on Social Network Analysis

Wen-Yuan Miao, Bing Fang*, Liu-Qing Meng

School of Management, Shanghai University, Shanghai, Shangda Rd 99, Shanghai, China
{lovedjango, melodyfang, mengliuqing}@i.shu.edu.cn

Received 21 March 2017; Revised 19 July 2017; Accepted 9 August 2017

Abstract. Retweet prediction plays an important role in understanding the diffusion of information on online social network sites. Current research mainly focuses on massive or individual retweet prediction. However, there exist few studies on retweet prediction within communities. To bridge this gap, this paper proposes a novel method to predict retweet behavior within communities based on social network analysis (SNA). First, a state-of-the-art community detection algorithm is applied to detect communities from the whole social graph. Second, the structure-related features of each detected community and their member are constructed using SNA. Third, a retweet prediction method within communities is proposed based on both structure related features and basic features commonly used in previous studies. An experiment is then performed to compare our proposed method with the baseline method, which uses basic features frequently used in previous studies. The results demonstrate the accuracy of adding network structural properties to retweet prediction within communities.

Keywords: community, information diffusion, retweet prediction, SNS, social network analysis

1 Introduction

Online social network sites (SNSs) have exploded in terms of scale and scope and have become one of the most important information diffusion channels over the last few years. Twitter, one of the most popular online SNSs, has gained a lot of research interest in the field of information diffusion, and retweet prediction has gradually become a popular research topic.

Current research mainly focuses on two aspects—massive retweet prediction and individual retweet prediction. Massive retweet prediction aims to predict the total number of retweets of each tweet, which can contribute to the understanding of information diffusion on an SNS [1] and for both commercial and political purposes [2-4]. In contrast, with the development of precision marketing, individual retweet prediction has emerged [5-6] for the purpose of personalized recommendations [5] and information filtering [7].

However, although there exist many studies on tweet popularity prediction, most of them ignore the community structure of an SNS. Community structure commonly exists on SNSs. People with similar interests, backgrounds, attitudes, and values form communities spontaneously [8]. Social psychological researchers have found that the actions of people belonging to the same community are homogeneous. For instance, the herd effect widely exists among the crowd [9]. According to social learning theory, people tend to learn social behavior by imitating and observing others, and this effect is also established in online SNSs, where people tend to follow their friends' actions on the social network [10-11].

In fact, communities play an important role in information diffusion. As illustrated in Fig. 1, the lifespan of information diffusion on an SNS can be divided into three stages: a tweet is retweeted by individuals first; then, it captures the imagination of various communities; and finally, it becomes a trending topic on the whole SNS [12-13]. Current studies mainly focus on the first and last stages of information diffusion on an SNS, i.e., individual retweet prediction and massive retweet prediction,

* Corresponding Author

respectively. Few studies concern the second stage of information diffusion on an SNS. However, tweets popular over an entire SNS are quite limited, and most of these tweets concern public issues, political events, or movie stars [14-15], while other information may just be popular within certain communities and generally cannot be detected by massive retweet prediction. Therefore, to deepen our understanding of information diffusion on the social network, the study of retweet prediction within communities is necessary and urgent.

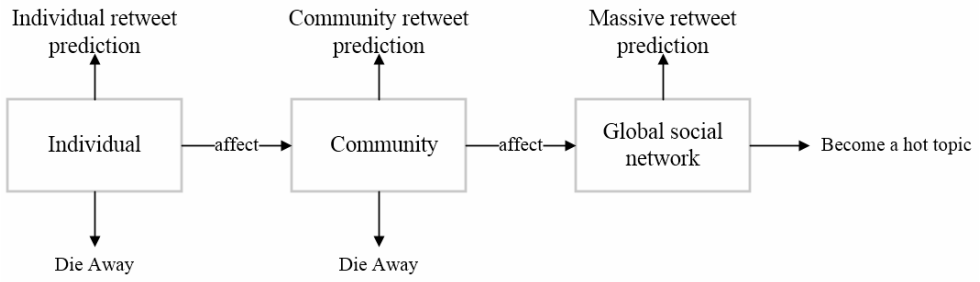

Fig. 1. Lifespan of information diffusion

The problem considered in this paper focuses on the second stage of information diffusion on an SNS, which can be called retweet prediction within communities. In particular, the research problem is formulated as a classification task and we train machine learning models that can automatically predict which tweet will become popular within a certain community.

Because the performance of machine learning models heavily relies on the quality of features passed to it [16]. To obtain better prediction results, social network analysis (SNA) was introduced to construct features from the perspective of network topology. Social psychology studies indicate that the activities of social network users and social influence between users are strongly associated with the structural properties of users [17-18]. SNA investigates the social structures of the network through network graph and graph theories [19], which models a social network in terms of nodes (users) and links (relationships) based on social network theory. Many studies have applied SNA to analyze the structural properties of an SNS. Chatfield and Brajawidagda [20] conducted a SNA of Twitter information flows among the Twitter followers of Central Disaster Agency; Ediger and Jiang [21] employed SNA to perform empirical studies at Twitter. Al-Sharawneh and Sinnappan [22] employed SNA to identify opinion leaders in Twitter during crisis situations. Despite the fact that few studies have introduced SNA or structural properties into the field of retweet prediction, the studies mentioned above have demonstrated the role of SNA in revealing users' SNS activities.

To predict retweets within communities, it is necessary to divide the whole social network graph into several separate communities. The technology used to find communities is called community detection. As a fundamental step for SNA, community detection is also currently a popular research topic [23]. The community detection algorithm used in this study is called the Louvain method, and it is able to detect communities effectively in large scale networks. Its details are discussed within the research framework.

In summary, the main technical achievements of our work are threefold:

(1) We design a novel framework to predict retweet behavior within communities.

(2) We introduce SNA techniques in the field of retweet prediction by detecting communities at SNS scale and construct structure-related features. Our findings indicate that it is feasible to predict retweet behavior while considering structure-related features.

(3) Our work reveals the properties of the second stage of information diffusion on SNSs, which builds a solid foundation for further study of the overall process of information diffusion on SNSs.

To the best of our knowledge, our work is the first to predict retweet behavior within communities and systematically apply SNA in the field of retweet prediction. This distinguishes our work from existing studies that are merely concerned with tweet popularity on the global network and which fail to systematically investigate the structural properties of an SNS. In this paper, we adopt a community detection algorithm, SNA, and machine learning methods to detect SNS communities, construct structure-related features and predict the popularity of tweets within communities. To test the performance of our proposed method, we formed a baseline method comprising basic features from the previous literature and compare it with our proposed method.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 introduces the overall framework of our research. The experimental steps are introduced in Section 4 and the results are analyzed in Section 5. Finally, Section 6 presents our conclusions.

2 Related Work

Current studies in retweet prediction can be divided into two types: massive retweet prediction and individual retweet prediction.

2.1 Massive Retweet Prediction

Studies on massive retweet prediction aim to predict the popularity of each tweet over the whole SNS. Morchid and Dufour [1] used principal component analysis (PCA) to analyze features including retweet times, hash tags, mentions, URLs for content features, and days (registration time), favorites, followers, followees, and status for user features, and predicted the tweet popularity based on these features using support vector machine (SVM) and NB. Lu [2] analyzed multimedia features and used PCA in feature selection to predict a tweet's global popularity. Deng and Ma [24] predicted the popularity of tweets concerning the conflicts between urban management officials (Chengguan in China) and the public using feature analysis and a back propagation neural network. They not only considered tweet-based features such as topic, mention, URL, and retweet times, but also constructed social-related features using the number of followers and followees and user degree of activity using the number of tweets the user has posted. This paper only focused on emergency issues and ignored structural properties. Ma and Sun [25] predicted the popularity of new emerging topics on Twitter within one day using naive Bayes, k-nearest neighbors, decision trees, support vector machines, and logistic regression (LR) models. All these studies predict tweet popularity through feature analysis and machine learning methods. Moreover, there also exist studies applying other technologies to predict tweet popularity. Bae and Ryu [26] predicted popularity through analyzing the similarity of tweet features. Wu and Shen [27] predicted the popularity of news from super nodes (news media nodes) using a general stochastic process-based approach. These studies all studied tweet popularity over the whole SNS, but did not include tweet popularity prediction within communities. Meanwhile, they did not systematically investigate the structural properties of an SNS.

2.2 Individual Retweet Prediction

Individual retweet prediction aims to predict whether a certain user will retweet certain tweets. To study this problem, both the features of tweet publishers and tweet receivers as well as their relationships and interactive information should be taken into consideration. Wang and Liu [28] predicted users' retweet action based on the features of released and accepted users and content, and they applied SVM to filter spam and established an LR model to complete prediction. Tang and Quan [29] performed an analysis of retweet features and improved LR for conducting the prediction. Zhou and Zhang [30] systematically examined the features of followee, follower, tweet, and interaction, and they included 52 features in total to perform the prediction.

Tang and Miao [5] made use of similarities between publishers and subscribers to train a machine learning model to conduct individual retweet prediction. Zhang and Tang [6] considered the influence of actions of a retweeter's friends and trained an LR model for individual retweet prediction. Lee and Mahmud [31] proposed a feature-based model that considers the content of tweets and social interactions. Most of these studies focused on the characteristics of receivers, publishers, and tweets while ignoring the connection between them. Meanwhile, although some of them introduced structure-related features (such as PageRank) in retweet prediction, they still failed to deeply investigate the structural properties between users, which can truthfully reflect users' social influence.

2.3 Main Differences between our Work and Previous Studies

Our work differs from the previous studies in two ways. First, the research problem is different from previous studies. While previous studies focus on the first (individual retweet prediction) and last

(massive retweet prediction) stages of information diffusion on SNS, our work only focuses on the second stage (retweet prediction within communities) of information diffusion on SNSs. Second, we introduce a new type of feature in retweet behavior prediction. Previous studies mainly use tweet-based and user-profile features for retweet behavior prediction. In addition to these two kinds of features commonly used in previous studies, using SNA techniques, we construct structure-related features to predict retweet behavior within communities.

3 Research Framework

According to the literature, most related studies combine machine learning methods with feature analysis to predict tweet popularity. The framework of our study, which is also based on this idea, is introduced in this section. The whole framework is shown in Fig. 2.

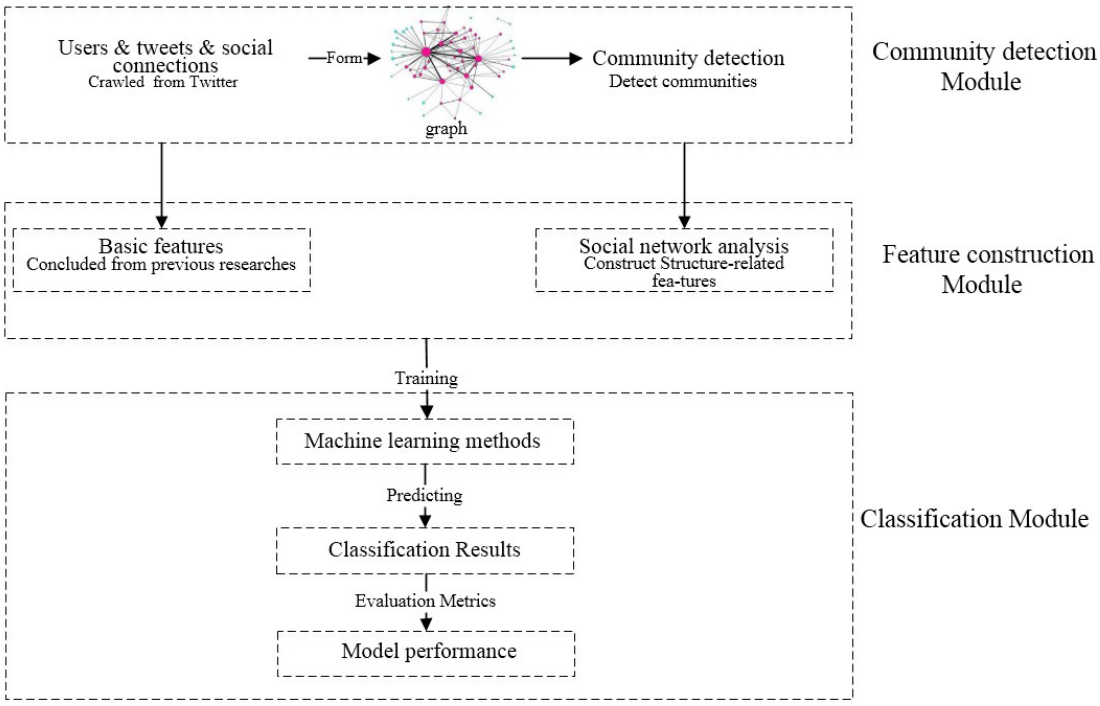

Fig. 2. Framework of this study

The framework of this study is divided into the following three modules: community detection, feature construction, and classification. In the community detection module, the whole social graph is divided into several separate communities through community detection technologies. In the feature construction module, both basic features used in previous studies and structure-related features produced by SNA are constructed. In the prediction module, both basic machine learning methods and ensemble methods are used to perform the classification with the constructed features. The details of each module are introduced below.

3.1 Community Detection

To predict tweet popularity within communities, it is necessary to first detect the communities. Community detection aims to partition a large network into several separate communities in which nodes from the same community are densely connected and nodes from different communities are sparsely connected. The algorithm of community detection used in this paper is called the Louvain method, which is a heuristic unsupervised method based on modularity optimization [8]. The Louvain method can detect communities quickly and effectively in huge networks with a time complexity of O(n). It uses modularity optimization to detect communities, which is demonstrated to be better suited to the social network partition problem because it is based on finding “community” structures in the network rather than a traditional graph partition [32]. The process of the Louvain method is as follows:

Step 1. Initialization: each point is divided into different communities.
Step 2. Each node is assigned to adjacent communities to calculate the highest increment of modularity. If the increment is positive, this assignment is accepted; otherwise, this assignment is not used.
Step 3. Repeat the process above until there is no positive increment in modularity.
Step 4. Each community is assigned to the adjacent community to produce the highest modularity increment. If the highest modularity increment is negative, the assignment is not used.
Step 5. Repeat Steps 2 to 4 until there is no change in the network modularity.

After all the steps above, each user is given a label that indicates the community he/she belongs to.

3.2 Feature Construction

Three kinds of features are constructed in our proposed method: tweet-based features, user-profile features, and structure-related features. Tweet-based features and user-profile features are often used in previous studies to reflect the properties of tweets and publishers, which are called as basic features; while structure-related features are rarely seen in the field of retweet prediction studies.

Basic features. We include basic features from previous studies, which include both tweet-based features and user-profile features.

Tweet-based Features. This group of features reflects the characteristics of tweets. They include whether tweets contain the character "#" (which indicates topic), whether tweets contain the character "@" (which means the tweet was tweeted at another user, i.e., @someone), whether tweets contain a URL (which indicates the multimedia), and the length the tweet [1, 2, 5, 24-26].

User-Profile Features. This group of features reflects the profile information of a tweet's original publishers. They are often used by researchers who study massive retweet prediction, which includes the number of followers, followees, and user collections as well as registration time [1, 2, 5, 6, 24, 26].

Structure-related features. There are two kinds of structure-related features that can truthfully reflect the social influence among users by calculating the structural properties of users and communities: user-structure features and community-structure features [33, 34]. Both can be constructed using SNA.

User-Structure Features. User–structure features can represent user structural properties that are closely related to the activities of users on SNSs [17, 18, 35, 36]. There are seven indicators in a user-structure feature, as follows:

(1) *Degree-Centrality*: users with higher degree-centrality are more influential within the community and their tweets are more likely to be retweeted. Its formula is

$$C'_D(p_k) = \frac{\sum_{i=1}^{n} a(p_i, p_k)}{N-1}. \quad \textbf{(1)}$$

where N is the number of total users in the community and $a(p_i, p_k) = 1$ if there is a connection between user i and user k; otherwise, $a(p_i, p_k) = 0$.

(2) *Betweenness-Centrality*: users with higher betweenness-centrality are seen as a "structural hole" in the community, which means they possess higher structural importance with respect to information diffusion. Its formula is

$$C'_B(p_k) = \frac{\sum_{i}^{n}\sum_{j}^{n} b_{ij}(p_k)}{N(N-1)/2}. \quad \textbf{(2)}$$

where N is the number of total users in the community and $\sum_{j}^{n} b_{ij}(p_k) = 1$ if the geodesics pass user k.

(3) *Closeness-Centrality*: users with higher closeness-centrality need fewer steps to transfer information to members within the community, and their information is more likely to be retweeted within the community. Its formula is

$$C'_C(p_k) = \left[\frac{\sum_{i=1}^{n} d(p_i, p_k)}{N-1} \right]^{-1} . \quad (3)$$

where N is the number of total users in the community and $d(p_i, p_k)$ is the length of shortest path between user i and user k.

(4) *Cluster-Score*: users with higher cluster-score have denser social links among their friends, and their information is more likely to be retweeted by their friends. Its formula is

$$ClusterScore = \frac{\#\text{Number of conntecion among A's friends}}{N(N-1)/2} . \quad (4)$$

where N is the number of total friends of user A.

(6) *Eigenvector-Centrality*: like PageRank, eigenvector-centrality is a measure of the user's structural importance. The eigenvector-centrality is defined as a weighted sum of the eigenvector centralities of adjacent vertices [34], and is calculated as follows:

$$\partial e(v_i) = \sum_{j=1}^{n} a_{ij} e(v_j) . \quad (5)$$

(7) *IsIncommunity*: This feature denotes whether the publisher is in the community. If the user is not in the community, he/she is added to the community temporarily to calculate the related features and then removed.

$$ISIncommunity = \begin{cases} 1, |\, \#\text{the tweet publisher is in the community}, \\ 0, |\, \#\text{the tweet publisher is not the community}. \end{cases} \quad (6)$$

Community-Structure Feature. The community-structure feature is closely related to information diffusion within the community. Previous studies have demonstrated that communication patterns between community members were strongly influenced by the topology of the community [38]. Community-structure features are represented by six indicators: four of them are degree-centralization [33], closeness-centralization [39], betweenness-centralization [40], and the clustering-coefficient [41], which show the amount of aggregation for degree-centrality, closeness-centrality, betweenness-centrality, and cluster-score of the community, respectively. Indicators for centralization have the following uniform formula.

$$C = \frac{\sum_{i=1}^{n} (C_{max} - C_i)}{max\left[\sum_{i=1}^{n} (C_{max} - C_i) \right]} . \quad (7)$$

The other two indicators are community size and density.

3.3 Classification

The classification module predicts which tweets will be popular within certain communities. In previous studies, most researchers identify popular tweets according to the times they were retweeted [1, 7, 24, 42]. In previous studies, researchers usually divided tweets into low retweeted tweets (LRTs) and massively retweeted tweets (MRTs) [1]. In this paper, we aim to predict tweet popularity within communities. Therefore, we formulate the prediction problem as a binary classification task for community LRTs (CLRTs) and community MRTs (CMRTs) based on the times these tweets were retweeted by the community

Based on the choices of previous studies, several binary machine learning methods were selected to perform the classification: SVM [1, 25], LR [5-6], and NB [1, 25] were chosen because they are frequently used in the field of retweet prediction.

In addition to the basic machine learning methods, the ensemble methods Adaboost (ADA) and

random forest (RF), are also included in this paper for comparison. The goal of ensemble methods is to combine several basic estimators with a learning algorithm to obtain a better generalizability than that of a single estimator. Ensemble methods can be categorized into averaging methods and boosting methods. Averaging methods aim to build several estimators independently and then average their predictions to reduce the variance and achieve better results. The chosen averaging method is RFs. Boosting methods aim to train basic estimators sequentially and produce a powerful ensemble by combining base estimators together. The chosen boosting method is ADA.

4 Experimental Study

An experimental study was designed to demonstrate the accuracy of our proposed retweet prediction method within communities. We implemented two algorithms for the experiments: our proposed method for systematically investigating structure-related features (PageRank excluded) and a baseline method that uses PageRank to represent structure-related features. The experiment was implemented using the following steps.

4.1 Data Collection

The dataset used for our experiments was collected from Twitter, the largest microblogging platform worldwide [43]. Because many accounts have been set to private by their owners or blocked for some reason, these accounts were removed and a social graph was crawled from twitter.com. The social graph contains users and their social links. Then, users' profile information and all of their posted tweets were also crawled. In this way, 22,017 users, 677,320 social links and 8,195,910 distinct tweets published by these users were obtained. The degree distribution of crawled users is shown in Fig. 3, which is a long tail distribution. All the data were collected by a web-spider programmed in Python.

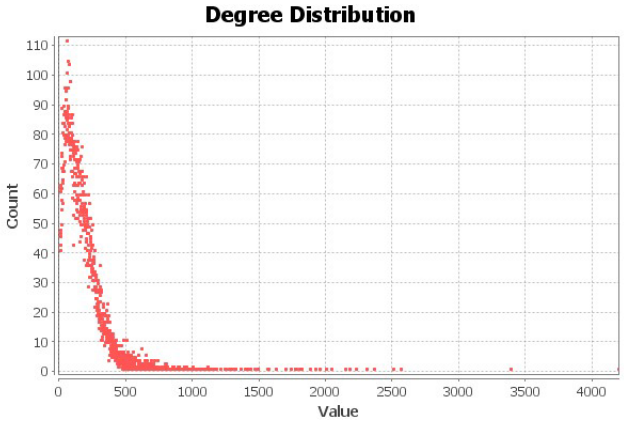

Fig. 3. Degree distribution

4.2 Community Detection

After obtaining data, the Louvain method was applied to detect communities from the crawled social graph. As a result, 77 communities were detected. The number of members in each community is shown in Fig. 4 and community statistics are shown in Table 1.

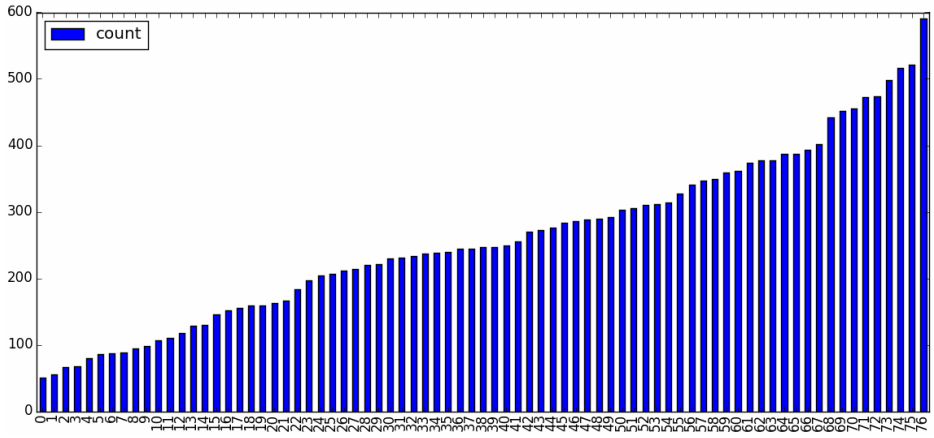

Fig. 4. Members of each community

Table 1. Statistical information of the communities

Count	Mean	Std	Max	Min
77	260	126	591	51

After dividing the whole social graph into 77 different communities, we employed a graph layout algorithm—forceatlas2 [44]—to visualize the ten largest communities in which nodes belonging to the same community are colored uniformly. This graph layout algorithm automatically clusters nodes in the same community and separates nodes belonging to different communities. The visualization result is shown in Fig. 5: the nodes of these ten communities account for 22.13% of all nodes in the whole social graph, while the edges of these ten communities account for 14.98%. As shown in Fig. 6, the inner structure of a community example is also visualized. In these two figures, nodes with high degree centrality are bigger than those with low degree centrality.

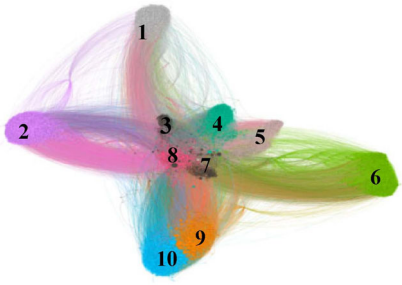

Fig. 5. Ten largest communities

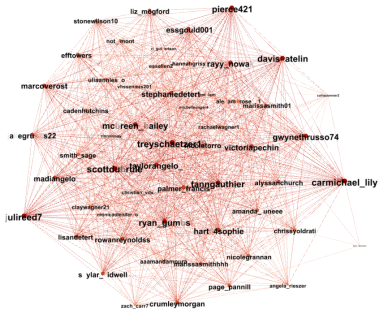

Fig. 6. Example of community structure

4.3 Feature Construction

Using the equations shown in Section 2, features were calculated for the collected data. Each feature was normalized using

$$x = \frac{x - min(x)}{max(x) - min(x)}. \tag{8}$$

All features used in both groups are summarized in Table 2.

Table 2. Summary of features

Category		
User-Profile	*Number of tweets*	Basic features
	Number of follower	
	Number of followee	
	Number of collection	
	Registration time	
Tweet-based Features	*IsTopic*	
	IsAlt	
	Len_of_twitter	
	IsUrl	
User-Structure Features	*PageRank*	Structure-related features
	DegreeCentrity	
	BetweennessCentrity	
	ClosenessCentrity	
	ClusterScore	
	EigenvectorCentrity	
	IsIncommunity	
Community-Structure Features	*Degree_Centralization*	
	Closeness_Centralization	
	Betweenness_Centralization	
	Clustering_Coefficient	
	Density	
	Size	

As discussed in the description of the framework, the retweet prediction features can be divided into basic features and structure-related features. As Table 3 shows both methods contain basic features. In addition, the proposed method contains all structure-related features based on SNA except PageRank, while the baseline method just makes use of PageRank as the structure-related feature.

Table 3. Training data and testing data

	Popular tweets	Unpopular tweets	Sum
Training data	8199	8199	16398
Testing data	2049	2049	4098
Sum	10248	10248	20496

4.4 Classification

We classified tweets into two types, CLRTs and CMRTs, according to their retweet times within communities. A tweet is labeled as popular within a community if it satisfies both the following conditions:

(1) It ranks in the top 1% most popular tweets in this community.

(2) It is retweeted more than five times in this community.

Because the number of unpopular tweets is far higher than the number of popular tweets, this will cause a serious data imbalance. To balance the data, the number of unpopular tweets, which were chosen randomly, was equal to the number of popular tweets. Machine learning methods introduced in research framework (including SVM, LR, NB, ADA, and RF) were used to perform classification. To obtain robust results, five-fold cross-validation was adopted. The amount of data in each fold is listed in Table 3.

5 Results and Discussion

5.1 Evaluation Metrics

Precision, recall, and the F1-measure were chosen to evaluate performance of the classification. The precision of retweet prediction refers to the total number of tweets that are correctly classified (true positives (TP) and true negatives (TN)) to total number of tweets. Because our research aims to identify popular tweets within communities, recall here refers to the ratio of the total number of tweets correctly classified as popular tweets (TPs) to the total number of popular tweets (the sum of TPs and FNs). The F1-measure comprehensively reflects precision and recall. The formulas of these three metrics are as follows.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{TN} + \boldsymbol{FN}}. \tag{9}$$

$$Recall = \frac{TP}{TP + FP}. \tag{10}$$

$$\boldsymbol{F1 - Measure} = \frac{2 \times \boldsymbol{precision} \times \boldsymbol{recall}}{\boldsymbol{precision} + \times \boldsymbol{recall}}. \tag{11}$$

5.2 Results and Discussion

The overall results of the five-fold experiment are shown below. Table 4 to Table 6 show the results of every fold and Fig. 7 to Fig. 9 show the average results of the five-fold experiment.

Table 4. Cross-validation results for precision

Fold	Group	NB	SVM	ADA	LR	RF
1	Proposed method	0.668	0.720	0.720	0.706	0.750
	Baseline method	0.642	0.660	0.712	0.703	0.729
2	Proposed method	0.648	0.713	0.721	0.699	0.758
	Baseline method	0.621	0.658	0.704	0.695	0.740
3	Proposed method	0.647	0.715	0.719	0.705	0.757
	Baseline method	0.640	0.657	0.725	0.696	0.763
4	Proposed method	0.621	0.704	0.722	0.696	0.757
	Baseline method	0.627	0.647	0.687	0.679	0.742
5	Proposed method	0.639	0.714	0.717	0.704	0.749
	Baseline method	0.633	0.650	0.703	0.691	0.732

Table 5. Cross-validation results for recall

Fold	Group	NB	SVM	ADA	LR	RF
1	Proposed method	0.599	0.816	0.821	0.755	0.817
	Baseline method	0.533	0.671	0.673	0.692	0.671
2	Proposed method	0.496	0.775	0.764	0.711	0.796
	Baseline method	0.442	0.659	0.615	0.652	0.642
3	Proposed method	0.475	0.739	0.733	0.700	0.765
	Baseline method	0.459	0.664	0.666	0.667	0.715
4	Proposed method	0.429	0.734	0.732	0.692	0.759
	Baseline method	0.402	0.649	0.559	0.604	0.626
5	Proposed method	0.487	0.743	0.736	0.687	0.745
	Baseline method	0.458	0.662	0.623	0.667	0.624

Table 6. Cross-validation results for F1-measure

Fold	Group	NB	SVM	ADA	LR	RF
1	Proposed method	0.632	0.765	0.767	0.730	0.782
	Baseline method	0.583	0.665	0.692	0.697	0.699
2	Proposed method	0.562	0.743	0.742	0.705	0.776
	Baseline method	0.516	0.659	0.657	0.673	0.688
3	Proposed method	0.548	0.727	0.726	0.702	0.761
	Baseline method	0.535	0.660	0.694	0.681	0.738
4	Proposed method	0.508	0.719	0.727	0.694	0.758
	Baseline method	0.490	0.648	0.616	0.639	0.679
5	Proposed method	0.552	0.728	0.727	0.695	0.747
	Baseline method	0.532	0.656	0.661	0.679	0.674

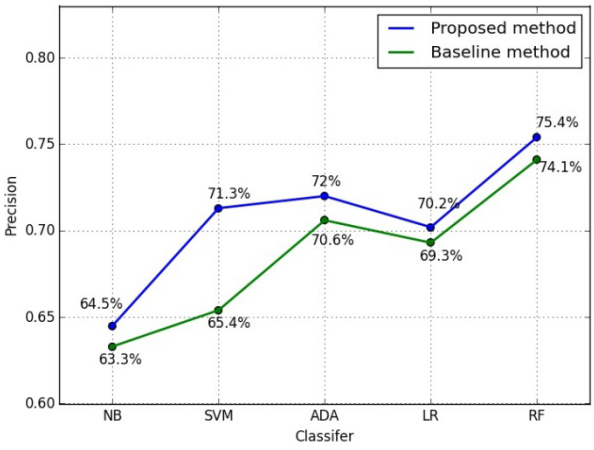

Fig. 7. Precision comparison

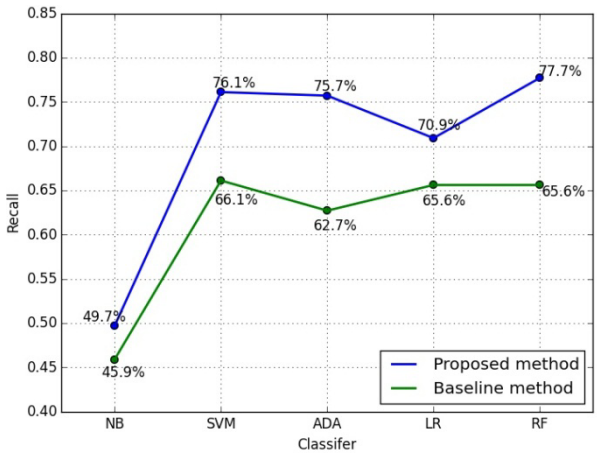

Fig. 8. Recall comparison

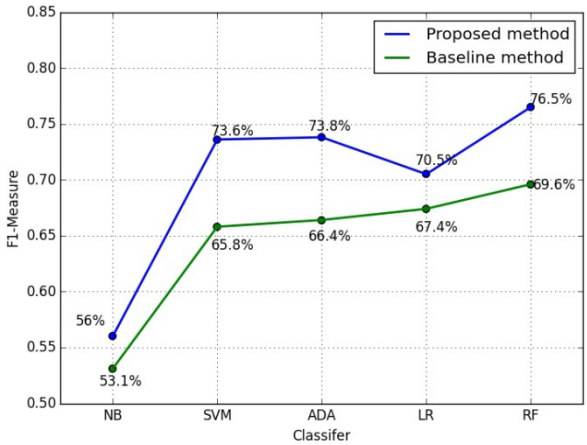

Fig. 9. F1-measure comparison

In Fig. 2, we find that no matter which classifier is used, the precision of our proposed method is always superior to the baseline method. The average increase is about 2.14%. The biggest improvement is from 65.4% to 71.3% for SVM, and the best precision of 75.4% is obtained by RF.

In Fig. 3, we also find that for all classifiers, the recall of our proposed method performs better than the baseline method. In addition, the recall of our proposed method is more improved than its precision. The average increase in precision is about 8.84%. The biggest improvement is from 65.6% to 77.7% for RF, and the best recall is 76.5%, also obtained by RF.

In Fig. 4, for all classifiers, our proposed method always performs better to the baseline method. The average increase in the F1-measure is about 5.6%. The biggest improvement is from 65.8% to 73.6% for

SVM, and the best precision is 76.5%, obtained by RF.

In summary, compared with the baseline method, our proposed method with the added structure-based features constructed by SNA performs better with respect to all three evaluation metrics. As for classifiers, we find that the best classification results are obtained by RF for all three evaluation metrics. In contrast, NB performs the worst for all three metrics. SVM and RF obtain the largest improvement when using our proposed method. Considering that the F1-measure comprehensively reflects recall and precision, the ensemble methods, RF and ADA, perform better than single classifiers.

6 Conclusions

In this paper, we proposed a novel research problem in the field of retweet prediction called retweet prediction within communities. To address this problem, we proposed a new method that uses SNA and machine learning methods. In our proposed method, a community detection algorithm is used to detect communities from the whole social graph, then the SNA is used to analyze each community and construct structure-related features. Finally both basic machine learning methods and ensemble methods are chosen to predict the popularity of each tweet within communities. To demonstrate the performance of our proposed method, an experiment was conducted with real data from Twitter. The experimental results show that our prediction method is more accurate than the previous method, which demonstrates that it is effective and necessary to apply structure-related features to the retweet prediction within communities task.

The main limitation of the paper is that content related features are not invovled in retweet prediction. In the future, the semantic analysis will be used to construct content related features. At the same time, based on the research results of this paper, we will further study the whole process of information flow on SNS. In particular, we will further investigate how information flows within one community and enters into other communities so as to reveal the mechanism of information flow on SNS from the perspective of communities.

Acknowledgements

This work is partially supported by Shanghai Natural Science Foundation of China (No. 16ZR1447100) and Shanghai Junior Faculty Cultivation Plan at Universities (No. N.37-0129-15-201). We thank Kim Moravec, PhD, from Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

References

[1] M. Morchid, R. Dufour, P.M. Bousquet, G. Linares, J.M. Torres-Moreno, Feature selection using principal component analysis for massive retweet detection, Pattern Recognition Letters 49(2014) 33-39.

[2] X. Lu, Z. Yu, B. Guo, X. Zhou, Predicting the content dissemination trends by repost behavior modeling in mobile social networks, Journal of Network & Computer Applications 42(3)(2014) 197-207.

[3] K.-H. Chu, J.B. Unger, J.-P. Allem, M. Pattarroyo, D. Soto, T.B. Cruz, H. Yang, L. Jiang, C.C. Yang, Diffusion of messages from an electronic cigarette brand to potential users through twitter, Plos One 10(12)(2015) e0145387.

[4] E. Kim, Y. Sung, H. Kang, Brand followers' retweeting behavior on twitter: how brand relationships influence brand electronic word-of-mouth, Computers in Human Behavior 37(2014) 18-25.

[5] X. Tang, Q. Miao, Y. Quan, J. Tang, K. Deng, Predicting individual retweet behavior by user similarity: a multi-task learning approach, Knowledge-Based Systems 89(C)(2015) 681-688.

[6] J. Zhang, J. Tang, J. Li, Y. Liu, C. Xing, Who influenced you? predicting retweet via social influence locality, Acm Transactions on Knowledge Discovery from Data 9(3)(2015) 1-26.

[7] W.M. Webberley, S.M. Allen, R.M. Whitaker, Retweeting beyond expectation: inferring interestingness in twitter, Computer Communications 73(2016) 229-235.

[8] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics Theory & Experiment 2008, abs/0803.0476.

[9] S.E. Asch, Opinions and social pressure, Scientific American 193(5)(1955) 31-35.

[10] L. Fan, Z. Lu, W. Wu, Y. Bi, A. Wang, B. Thuraisingham, An individual-based model of information diffusion combining friends' influence, Journal of Combinatorial Optimization 28(3)(2014) 529-539.

[11] Y.D. Seo, Y.G. Kim , E. Lee, D.K. Baik, Personalized recommender system based on friendship strength in social network services, Expert Systems with Applications 69(2017) 135-148.

[12] S. Glezos, Virtuous networks: machiavelli, speed and global social movements, International Politics 53(4)(2016) 534-554.

[13] K. Yang, X.L. Liu, J.H. Lin, X. Ceng, Q. Guo, J.G. Liu, The evolution of social networks constructed by "Ice Bucket Challenge," Complex Systems and Complexity Science 13(2016) 90.

[14] A. Bruns, T. Highfield, J. Burgess, The Arab spring and social media audiences English and Arabic twitter users and their networks, American Behavioral Scientist 57(7)(2013) 871-898.

[15] S. Park, J. Lee, S. Ryu, K.S. Hahn, The network of celebrity politics: political implications of celebrity following on twitter, The ANNALS of the American Academy of Political and Social Science 659(1)(2015) 246-258.

[16] B. Agarwal, N. Mittal, Semantic feature clustering for sentiment analysis of English reviews, IETE Journal of Research 60(6)(2014) 414-422.

[17] Z. Katona, P.P. Zubcsek, M. Sarvary, Network effects and personal influences: diffusion of an online social network, Journal of Marketing Research 48(48)(2011) 425-443.

[18] A. Klein, H. Ahlf, V. Sharma, Social activity and structural centrality in online social networks, Telematics & Informatics 32(2)(2015) 321-332.

[19] E. Ozyar, S. Gurdalli, Social network analysis: a powerful strategy, also for the information sciences, Journal of Information Science 28(6)(2002) 441-453.

[20] A.T. Chatfield, U. Brajawidagda, Twitter tsunami early warning network: a social network analysis of twitter information flows, in: Proc. Australasian Conference on Information Systems, 2012.

[21] D. Ediger, K. Jiang, J. Riedy, R. Farber, W.N. Reynolds, Massive social network analysis: mining twitter for social good, in: Proc. International Conference on Parallel Processing, 2010.

[22] J. Al-Sharawneh, S. Jebrin, M.A. Williams, Credibility-based twitter social network analysis, in: Proc. Web Technologies and Applications, 2013.

[23] M. Ozer, N. Kim, H. Davulcu, Community detection in political twitter networks using nonnegative matrix factorization methods, in: Proc. Ieee/acm International Conference on Advances in Social Networks Analysis and Mining, 2016.

[24] Q. Deng, Y. Ma, Y. Liu, H. Zhang, Prediction of retweet counts by a back propagation neural network, Journal of Tsinghua University (Science and Technology) 55(12)(2016) 1342-1347.

[25] Z. Ma, A. Sun, G. Cong, On predicting the popularity of newly emerging hashtags in twitter, Journal of the Association for Information Science and Technology 64(7)(2013) 1399-1410.

[26] Y. Bae, P.M. Ryu, H.K. Kim, Predicting the lifespan and retweet times of tweets based on multiple feature analysis, Etri Journal 36(3)(2014) 418-428.

[27] B. Wu, H. Shen, Analyzing and predicting news popularity on twitter, International Journal of Information Management

35(6)(2015) 702-711.

[28] Z. Wang, K. Liu, Z. Zheng, Prediction retweeting of microblog based on logistic regression model, Journal of Chinese Computer Systems 37(8)(2016) 1651-1655.

[29] X. Tang, Y. Quan, J. Song, K. Deng, H. Zhu, Q. Miao, Novel algorithm for predicting personalized retweet behavior, Journal of Xidian University, 4(2016) 57-62+68.

[30] J. Zhou, Z. Zhang, B. Wang, Y. Zhang, Y. Yan, Predicting who will retweet or not in microblogs network, in: Proc. Chinese National Conference on Social Media Processing, 2015.

[31] K. Lee, J. Mahmud, J. Chen, M. Zhou, J. Nichols, Who will retweet this? detecting strangers from twitter to retweet information, Acm Transactions on Intelligent Systems & Technology 6(3)(2015) 1-25.

[32] J.M. Pujol, V. Erramilli, P. Rodriguez, Divide and conquer: partitioning online social networks, in: Proc. Computer Science, 2012.

[33] J.M. Mccullough, E. Eisencohen, S.B. Salas, Partnership capacity for community health improvement plan implementation: findings from a social network analysis, BMC Public Health16(1)(2016) 566.

[34] A. Nikakhtar, S.A. Abbasian-Hosseini, H. Gazula, S.M. Hsiang, Social network based sensitivity analysis for patient flow using computer simulation, Computers & Industrial Engineering 88(2015) 264-272.

[35] S.P. Borgatti, A. Mehra, D.J. Brass, G. Labianca, Network analysis in the social sciences, Science 323(5916)(2009) 892.

[36] A. Rezvanian, M.R. Meybodi, Stochastic graph as a model for social networks, Computers in Human Behavior 64(2016) 621-640.

[37] Y. Wang, L. Bi, S. Lin, M. Li, H. Shi, A complex network-based importance measure for mechatronics systems, Physica A Statistical Mechanics & Its Applications 466(2016) 180-198.

[38] A.J. Morales, J. Borondo, J.C. Losada, R.M. Benito, Efficiency of human activity on information spreading on twitter, Social Networks 39(1)(2014) 1-11.

[39] M.J. Kim, H. Ahn, A theoretical framework for closeness centralization measurements in a workflow-supported organization, Ksii Transactions on Internet & Information Systems 9(9)(2015) 3611-3634.

[40] M.W. Schoen, S. Morelandrussell, K. Prewitt, B.J. Carothers, Social network analysis of public health programs to measure partnership, Social Science & Medicine 123(2014) 90-95.

[41] S.J. Hardiman, L. Katzir, Estimating clustering coefficients and size of social networks via random walk, in: Proc. International Conference on World Wide Web, 2013.

[42] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in twitter, in: Proc. International Conference on World Wide Web, 2011.

[43] J.M. Cotelo, F.L. Cruz, F. Enriquez, J. Troyano, Tweet categorization by combining content and structural knowledge, Information Fusion 31(C)(2016) 54-64.

[44] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, Plos One 9(6)(2013) e98679.