Predicting User Retweeting Behavior in Social Networks With a Novel Ensemble Learning Approach

LONG CHEN and HUIFANG DENG
Department of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
Corresponding author: Long Chen (longkcscn@outlook.com)

ABSTRACT
Information sharing through online social networks has become a main mechanism by which people share information with their friends through retweeting behaviors, which may result in a variety of information diffusion cascades on social media such as Facebook, Twitter, and Weibo. Predicting user retweeting behavior in those social networks is extremely challenging. To complete the prediction task, identifying factors that affect retweeting behavior and constructing efficient model are necessary. In this paper, we study heterogeneous relation networks by considering various social interactions, which reflect how a particular retweeting action is affected by the social behaviors performed by the sender and the receiver of the retweet. We then generate various features from our identified factors belonging to three dimensions — content semantics, user diffusion behavior, and network structure. Moreover, we cast our prediction problem as an ensemble learning problem and propose a novel ensemble learning approach to solve the problem. Combing the generated features and the novel ensemble learning approach, we then propose a model named Retweeting Behavior on Multiple Heterogeneous Diffusion Relation Networks (RBMHDRN) to predict user retweeting behavior in social networks. Experiments on a real dataset extracted from a social network site Weibo demonstrate the effectiveness of our proposed model, indicating that our generated features and proposed approach can significantly improve the performance of predicting user retweeting behavior occurring in the process of information diffusion in social networks.

INDEX TERMS
Information diffusion, retweet prediction, ensemble learning, online social networks, user diffusion behavior.

I. INTRODUCTION
Rumor, disease, and innovation, each of them can be regarded as a piece of information that spreads over the edges of social networks. With the occurrence of social media sites like Facebook, Twitter and Weibo, various categories of information have been propagating in online social networks, which become a significant mechanism of information diffusion.

In an online social network, when a user is exposed to a piece of information posted by others, he/she can read the information and choose to retweet it, thus to make his/her friends be exposed to the information as well. Then, his/her friends also can read the information and choose to retweet it, thus to make their respective sets of friends be exposed to the information as well. The above process repeats and may result in an information diffusion cascade. Compared to classical information diffusion that occurs in real world, the information diffusion where people adopt online social media to spread their texts, images and videos, in essence, is similar, except that the range is wider, the speed is higher, and the network structure is more explicit.

Moreover, online social media has become a powerful marketing force as it can benefit promoting the awareness of a brand, predicting the popularity trend of an event [1], and increasing product sales [2]. Research interest has continually grown on information spreading dynamics in social networks [3], while user retweeting behavior is usually considered to be the main essence of the viral aspect of information spreading in social media [4]. To capture the dynamics of information diffusion occurring in online social networks, it is necessary to analyze the information diffusion data to obtain the factors that affects the diffusion process and
construct a model to predict whether the retweeting behaviors of users will happen in the future. Consequently, modeling and predicting user retweeting behavior in online social networks has become an important research problem.

A growing body of researches has been developed on information diffusion in online social networks, such as identifying important spreaders by following real diffusion dynamics in social media [5], measuring the structural virality of information diffusion in online social networks [6], and predicting popular information propagated in a famous social media - Twitter [7]. However, most of the researches consider that there is only one network can influence information diffusion in social networks, and the network is generally based on the follow relationships between users.

In this paper, we consider that information diffusion in online social networks is influenced by multiple heterogeneous diffusion relation networks, which are constructed on various social interactions between users. Information content and user diffusion behaviors are also considered by us when predicting user retweeting behavior that occurs in the process of information diffusion in social networks. In addition, we cast the prediction problem as an ensemble learning problem. The main contributions of this paper can be summarized as follows.

1) We define and construct multiple heterogeneous diffusion relation networks, based on various social interactions between users. Then, we generate features from our identified properties belonging to three dimensions: user diffusion behavior, network structure, and content semantics. Moreover, we assume that user retweeting behavior is mainly influenced by those multiple heterogeneous diffusion relation networks.

2) We consider the prediction problem as an ensemble learning problem and propose a novel ensemble learning approach named Parallel Adaboost on Logistic Regression with a Single Feature (PALRSF) to solve the problem. Combing the generated features and the proposed novel ensemble learning approach, a novel model named Retweeting Behavior on Multiple Heterogeneous Diffusion Relation Networks (RBMHDRN) is proposed to predict user retweeting behaviors in online social networks.

3) Various experiments are conducted on a real world dataset extracted from Weibo. First, we compare our proposed prediction model RBMHDRN with two simple baselines, and one of the state-of-the-art models – LRC-BQ to validate the effectiveness of the proposed model. We also compare our proposed model with other three state-of-the-art models. Then, we compare our proposed novel ensemble learning approach PALRSF with three classic methods, especially with two of them when using different sizes of the training set. Moreover, we evaluate the speedup performance of RBMHDRN as well.

The paper is organized as follows. In Section II we introduce some related work. Section III shows the analysis of information diffusion process in online social networks and the definition of the problem of predicting user retweeting behavior. In Section IV, we describe various features generated from our identified factors belonging to three dimensions, and present our proposed prediction model that combines those generated features with a novel ensemble learning approach to complete the prediction task. Section V, the details of the dataset and experiments are described, as well as some analysis of the experimental results. In Section VI we conclude and discuss some future work of this study.

II. RELATED WORK

Many papers have researched the theoretical models of information diffusion in networks, and utilized those models to simulate the diffusion dynamics. The model SIR [8] was developed to model the propagation of virus, where S represents the “susceptible” state, I represents the “infected” state, and R represents the “recovered” state. Zhan et al. in [9] analyzed the process of virus propagation both in real world and on the Internet, then constructed a nonlinear model based on the SIS model to explore the coupling effect. Kwan et al. integrated newly diagnosed HIV patients’ phylogenetic, clinical and behavioral data, and incorporated an information diffusion model to analyze transmission dynamics [10].

In addition, the Linear Threshold Model [11] was used to study collective behavior at first, and then was applied to model information diffusion as well. For example, Galuba et al. [12] proposed a method to predict the diffusion graph using the Linear Threshold Model, with the beginning of the diffusion process is observed. The Independent Cascade Model [13] considers the dynamic cascade of the diffusion process, in which each edge of the network is assigned a uniform probability, as a node may activate a neighbor through the edge between them. Saito et al. incorporated synchronous time-delay parameter to the Independent Cascade Model, and they proceeded iteratively along a continuous time axis [14]. Kempe et al. adopt classical models including the Linear Threshold Model and the Independent Cascade Model to address the optimization problem of selecting the most influential nodes [15].

While other researches have analyzed empirically observed information flow in social networks, and consider effective methods to predict the properties of real-world information diffusion cascades. Meanwhile, the researches that aim to build a prediction model for the propagation of information generally adopt machine learning techniques [16] as methods. Considering that a user might be exposed to a microblog several times by various followees before he/she decides to retweet it, Hu et al. performed an empirical study to understand the selectivity of retweeting behaviors, and proposed an individual interaction model to infer which followee user will choose [17]. Shi et al. in [18] developed an integrated conceptual framework to investigate the mechanism of individual retweeting behavior, and found that social

VOLUME 8, 2020
148251
strength and topical relevance with the receiver are the most important factors. In [19], an event detection method taking advantage of retweeting behavior was proposed by Chen et al. to handle the events evolution. They proposed a topic model RL-LDA capturing the social media information over hashtag, location, textual, and retweeting behavior to handle complex event. Li and Liu [20] explored the patterns of retweeting behavior of influential microblogs and established a classification model based on social-influential, topical, and temporal factors to predict the temporal class of an original microblog’s retweeting time series. Romero et al. analyzed the ways in which hashtags spread over the social network of Twitter using statistical methods to acquire the exposure-curve, and found significant variation in the ways that widely used hashtags on different topics spread [21]. Szabo et al. presented a method to predict the long time popularity of Digg stories and Youtube videos from early measurements of access [22].

Moreover, Wang et al. [23] proposed the concept of the value strength, social strength and a time-varying graph (TVG)-based mobility model, based on which a forwarding nodes’ selection scheme and a socially-aware information transmission mechanism were presented, which can significantly improve the information propagation efficiency and information coverage ratio of mobile networks. The interactions among different contagions when they spread through the network were studied by Myers et al. in [24], the authors developed a statistical model incorporating competition and cooperation of different contagious in information diffusion. Liu et al. in [25] incorporated homogeneity trend into a modified competitive rumor model with generalized population preference and examined how competitive information diffuses and evolves in social networks. Cheng et al. [26] considered the prediction of cascade growth as binomial classification and found that for a cascade with $k$ first retweets, predicting the final size whether will reach the median size of all the cascades with at least $k$ retweets, is equivalent to if the cascade will become double times on size. In addition, they tried a variety of learning methods, including random forests, SVM, and linear regression, to realize the prediction.

The researches most relevant to our work are those that aim to predict the dynamics of information diffusion in online social networks. A novel information diffusion model was proposed by Li et al. in [27], and the proposed model treats users as intelligent and relational agents and calculates the corresponding payoffs of them to predict whether the behaviors of users will happen. Chen et al. [28] considered users’ retweet behaviors, focusing on whether users with a certain emotional status will retweet the tweet and proposed a retweeting prediction framework. Zhang et al. in [29] defined the notion of social influence locality and presented the corresponding instantiation functions based on pairwise influence and structure influence, then proposed a method named LRC-BQ combing the influence locality function and basic features including personal attributes, instantaneity and topic propensity to predict retweeting behaviors. Zhang et al. [30] proposed a novel method using hierarchical Dirichlet process to combine structural, textual, and temporal information to predict retweet behavior. Besides, Jiang et al. [31] proposed an evolutionary game theoretic framework to model the dynamic process of information diffusion in social networks and derive the dynamics in various networks, and they found that the dynamics of information diffusion over those networks are scale-free and the same with each other when the network scale is sufficiently large. Simulations performed on both synthetic networks and real-world Facebook networks showed that the proposed model could fit and predict the information diffusion over real social networks well. Moreover, they also conducted experiments on real-world information spreading dataset of Memetracker, which showed that the proposed framework is effective and practical in modeling the user forwarding behaviors in social networks [32]. Zhu et al. presented a novel solution to model the retweeting behavior over user groups and developed a system named GruBa to extract user-based features, cluster users into groups, and model retweeting behavior [33].

Most of the existing methods of predicting user retweeting behavior in online social networks assume that, only one single network, only one type of user behavior, or only one kind of user can influence the process of information diffusion in online social networks. This may lead to a common shortcoming: several features that affect the process of information diffusion significantly are not considered. However, our proposed method assumes that, multiple heterogeneous diffusion relation networks, various types of user diffusion behaviors, and various kinds of users can influence the process of information diffusion, which is consistent with the reality and can take more features that are important into account to research the information diffusion occurring in online social networks.

III. PRELIMINARY AND PROBLEM FORMULATION

To research the prediction problem of whether a user will retweet a piece of information or not, a real dataset of information diffusion cascades is necessary. In this paper, we extracted the dataset from a real-world online social network – Weibo, in which the spreading information is microblog.

A. ANALYSIS OF INFORMATION DIFFUSION PROCESS

Based on the observed data, we analyze the practical process of information diffusion in social networks. Once a piece of information is exposed to a user, called target user, he/she decides whether to spread the information by retweeting behavior. Meanwhile the retweeting behaviors of the target user’s neighbors have significant influence on the retweeting behavior of the target user, as many researches have proved [34]. Generally, the information is initially posted by a user, called initial user. Then the information is exposed to a target user via another user, called parent user. In addition, there is a category of social relationship between the target user and the parent user. Most prior researches treat
following relationship as the social relationship that supports information propagation between users. However, following relationship is not sufficient to support information diffusion, as it is not equivalent to the existence of historical information diffusion between users.

In this paper, we extract different categories of social relationships between users from various social interactions. We treat the social relationship extracted from retweeting interactions between users as the social relationship mainly supporting information diffusion, because a retweeting interaction between users means the existence of historical information propagation between them, which is more possible to trigger future information propagation between the same users.

By analyzing the observations of information diffusion process in Weibo, we find that we can extract three categories of social relationships between users from historic social interactions. First, we extract the indirect retweeting relationships with the following method: If user A retweets a microblog initially posted by user B, then A has an indirect retweeting relationship to B. Second, we extract the direct retweeting relationships: If user A directly retweets a microblog posted by user C, then A has a direct retweeting relationship to C. Third, we extract the mentioning relationships: If user A mentions another user D, then A has a mentioning relationship to C.

### B. DIFFUSION RELATION NETWORKS

A heterogeneous network or heterogeneous information network [35] is defined as a network in which the number of the types of objects is more than one or the number of the types of links is more than one. Heterogeneous network provides new perspectives to manage networked data and introduces new challenges for many data mining tasks. For example, Wang et al. [36] proposed a model to learn relevance for online targeting by analyzing of the context of heterogeneous networks. Tang et al. in [37] developed another application named PatentMiner 6 to realize topic-driven analyzing and mining of heterogeneous patent networks.

Although many data mining tasks have been exploited in heterogeneous network, there are still some challenging research issues. In a heterogeneous network, one object can refer to several entities in real world, and several objects can refer to one entity, e.g., bibliography data can contain duplication of name [38]. Relationships between objects may not be explicitly known or not complete, e.g., the relationships between advisors and advisees in the DBLP network [39]. Moreover, a heterogeneous network may contain attribute values representing important information on edges, then the effect of those attribute values on the weighted heterogeneous network is necessary to be considered [40].

In this paper, with the extracted various categories of social relationships between users, we can construct a heterogeneous network that has only one type of objects — users and various types of social relationships between those objects. Furthermore, we can split this heterogeneous network into multiple heterogeneous diffusion relation networks, each of which has the same type of objects — users, but a different type of social relationships between those objects.

For the convenience of later discussion, we give the main symbols that will be used throughout the paper and their corresponding detailed descriptions in Table 1.

**Definition 1 (Diffusion Relation Network):** A diffusion relation network is a network constructed on some category of social relationship between users that can help information to diffuse and can be represented as $H = (V, R)$, where $V$ represents the set of users and $R$ represents the set of edges corresponding to this category of social relationships between users.

According to the above three categories of social relationships between users, we can construct three
corresponding heterogeneous diffusion relation networks, including the direct retweeting network \( H_d = (V, R_d) \), the indirect retweeting network \( H_r = (V, R_i) \), and the mentioning network \( H_m = (V, R_m) \). Note that, \( R_d \) is a set of edges: An edge \((u, v) \in R_d\) represents a direct retweeting relationship between user \( u \) and user \( v \). \( R_i \) is another set of edges: An edge \((u, v) \in R_i\) represents an indirect retweeting relationship between user \( u \) and user \( v \). And \( R_m \) is another set of edges: An edge \((u, v) \in R_m\) represents a mentioning relationship between user \( u \) and user \( v \).

**Definition 2 (User Diffusion Behavior):** A user diffusion behavior is a kind of user behavior that aims to diffuse information and can be represented as a triple \( B = (u, m, a) \), which can be interpreted as that, user \( u \) performs an action \( a \) through microblog \( m \) to spread information to other users.

In this paper, we identify three types of user diffusion behaviors, including user posting behavior \( B_p = (u, m, a_p) \), user retweeting behavior \( B_r = (u, m, a_r) \), and user mentioning behavior \( B_m = (u, m, a_m) \), where \( a_p, a_r, \) and \( a_m \) represent the action of posting, the action of retweeting, and the action of mentioning, respectively.

Moreover, we assume that information diffusion is also influenced by different types of user diffusion behavior, while most prior researches usually consider only one type of user diffusion behavior – propagation behavior.

**C. PROBLEM STATEMENT**

Information diffusion can be regarded as a process in which a piece of information spreads between users, in the environment of networks. Our aim in this paper is to study how to predict user retweeting behavior occurring in the process of information diffusion in social networks, which needs to explore various factors prompting the retweeting behaviors of users.

**Definition 3 (Predicting Retweeting Behavior):** We define a prediction task that we intend to predict whether a user will retweet a given microblog after one of his/her parent users has retweeted or initially posted the microblog.

Fig. 1 illustrates the problem of predicting retweeting behavior with an example: an initial user \( v_0 \) posts a public microblog, seen by other three users \( v_1, v_2, \) and \( v_3 \), as \( v_0 \) is one of their common parent users. Then, users \( v_2 \) and \( v_3 \) decide to retweet the microblog, while user \( v_1 \) has not decided to retweet the microblog. This way the microblog propagates over the edges of the network, and a user is exposed to the microblog via his/her parent users who have retweeted or initially posted the microblog. Each user who has retweeted the microblog is represented as a blue node, which is connected to its retweeting source node through a solid edge. Each user who is exposed to the microblog and has not retweeted the microblog is represented as a green node, which is connected to its source node through a dashed edge. Then, we can predict the retweeting behaviors of \( v_1, v_5, \) and \( v_9 \), as each of them is exposed to the microblog via his/her parent users, and may decide to retweet the microblog in the future.

**IV. PROPOSED MODEL**

There are several existing approaches have been adopted to solve the prediction of information diffusion. Some researchers use statistical methods to acquire the exposure-curve [26], and demonstrate its effectiveness with significant performance. Numbers of researchers incorporate machine learning methods [16] used in classification tasks and get comparative results as well. Some researchers adopt classical models like the Linear Threshold Model [15], where the prediction method becomes assigning a threshold to each node. In this paper, we decide to consider the prediction problem as an ensemble learning problem and propose a prediction model which adopts a novel proposed ensemble learning approach to predict user retweeting behavior. Moreover, the proposed model used to predict user retweeting behavior relies on the assumption that a target user will decide to retweet a piece of information only after one of his/her parent users has retweeted or initially posted the information.

**A. FACTORS**

To predict user retweeting behavior in a social network, it is important to identify the factors affecting the behavior at first. We identify the factors belonging to three dimensions – content semantics, network structure, and user diffusion behavior, then generate the corresponding features to prepare for predicting user retweeting behavior.
Most prior researches extract the corresponding features of only one network, only one kind of user, or only one type of user diffusion behavior. However, in this paper, we generate the corresponding features of multiple heterogeneous diffusion relation networks, various kinds of users, and various kinds of user diffusion behaviors.

Moreover, when quantifying some of the corresponding features, it is necessary to normalize the values of those features, and the normalization is shown in the following formula:

\[
\text{Normalization:} \quad \text{norm}(f_i) = \frac{f_i - \min(f)}{\max(f) - \min(f)} \quad (1)
\]

where \(f\) presents a feature, \(f_i\) is the corresponding value of this feature for \(i\)-th instance, \(\max(f)\) is the maximum value of this feature for all the instances, and \(\min(f)\) is the minimum value of this feature for all the instances.

1) DIMENSION OF CONTENT SEMANTICS

To capture the properties of content semantics, we analyze the content of all the microblogs as follows: First, we extract keywords from the content of all the microblogs. Then, we calculate the corresponding weights of the extracted keywords using the TF-IDF score [41] and obtain the topic distributions of all the microblogs using the Latent Dirichlet Allocation (LDA) model [42]. As a user has posted microblogs according to his/her interest, the content of these microblogs can be used to represent the interest of the user. Based on the above works, three types of semantic similarities between words of a user can be gathered from the keywords of all the microblogs posted by him/her.

Keyword Similarity: With the keywords extracted from the content of microblogs, we consider one type of semantic similarity, which indicates how many keywords a user employed are the same with the keywords in the content of a microblog, called keyword similarity. The similarity can be calculated as follows:

\[
\text{Keyword Similarity:} \quad s_{\text{key}}(u, m) = \frac{|C_u \cap C_m|}{|C_m|} \quad (2)
\]

where \(C_u\) and \(C_m\) denote the keywords employed by user \(u\) and microblog \(m\), respectively, and the keywords of a microblog can be directly extracted from its content, the keywords of a user can be gathered from the keywords of all the microblogs posted by him/her.

TF-IDF Similarity: As the keywords of microblogs are obtained, a microblog can be represented by the TF-IDF representation, i.e., the microblog can be represented as an \(n\)-dimensional vector and each dimension of the vector represents a distinct keyword term, whose weight can be calculated using the TF-IDF score [41].

Moreover, the TF-IDF representation of a user can be obtained by adding up the TF-IDF representations of all the microblogs posted by the user. Then, the TF-IDF similarity between a user and a microblog can be calculated using cosine similarity.

\[
\text{LDA Similarity:} \quad \text{After the content of microblog } m \text{ is handled by the LDA model, the topic distribution of the microblog can be represented as } P_m = (p_{m1}, p_{m2}, \ldots, p_{mT}), \text{ where } p_{mi} \text{ is the probability of topic } i \text{ in the content of } m \text{ and } T \text{ is the number of topics.}
\]

Moreover, the topic distribution of a user can be obtained by adding up the topic distributions of all the microblogs posted by the user.

Then, the LDA similarity between user \(u\) and microblog \(m\) can be calculated as following:

\[
\text{LDA Similarity:} \quad s_{\text{lda}}(u, m) = \frac{1}{1 + T\sqrt{\sum_{i=1}^{T}(p_u^i - p_m^i)^2}} \quad (3)
\]

2) DIMENSION OF NETWORK STRUCTURE

Network is essential for information diffusion because it provides the medium through which the information propagates. As described in Section III-B, we consider three heterogeneous diffusion relation networks, assuming that the process of information diffusion is influenced by multiple heterogeneous diffusion relation networks, while the existing researches generally take only one relation network into consideration. Then, we capture the structural properties of those multiple heterogeneous diffusion relation networks as follows.

Relationship Edge: A relationship edge represents the existence of some kind of social relationship between users. As the edges that represent social relationships can be used to construct a network by which information spreads, relationship edges may contribute to user retweeting behaviors.

\[
\text{Relationship Edge:} \quad t_{u,v}^R = \begin{cases} 
1, & \text{if } (u, v) \in R \\
0, & \text{otherwise} 
\end{cases} \quad (4)
\]

Note that we have extracted three categories of social relationships between users, thus \(R\) can be \(R_d, R_i, \text{ or } R_m\).  

Degree: Degree, which refers to the number of connections of a node in a network, is used widely as a structural measure of user influence in social networks. On each diffusion relation network, we calculate the in-degree values of users and normalize those values.

K-Core: K-core [43], also called k-shell, describes the location of a node in a network. On each diffusion relation network, we calculate the k-core values of users and normalize those values.

Pagerank: We adopt PageRank [44] to evaluate the structural importance of users in each diffusion relation network. For user \(u\) of diffusion relation network \(H = (V, R)\), we can calculate his/her PageRank value as follows:

\[
\text{Pagerank:} \quad r_u^R = \sum_{v: (v, u) \in R} r_v^R (q \cdot t_{v,u}^R + \frac{1 - q}{|V|}) \quad (5)
\]

where \(q\) represents the damping factor, \(t_{v,u}^R\) represents the transition probability over edge \((v, u)\) and the calculation is...
shown in the following formula.

$$t^R_{v,u} = \frac{E^R_{v,u}}{\sum_{E(v,i) \in R} E^R_{v,i}}$$

(6)

where $E^R_{v,u}$ is the number of times the social relationship between user $v$ and user $u$ has appeared. We calculate the PageRank values of users and normalize those values with (1).

3) DIMENSION OF USER DIFFUSION BEHAVIOR

As described in Section III-B, we assume that several types of user diffusion behaviors are influential to information diffusion in social networks, and identify three types of user diffusion behaviors in Weibo. For each type of user diffusion behavior, we mainly focus on the frequency of this behavior performed by user to capture the properties of user diffusion behavior as follows:

**Behavior Activity:** The behavior activity of a user expresses the relative activity of the user on performing some type of user diffusion behavior and can be calculated by the following formula:

$$A^B_u = \sum_{v \in V, v \neq u} \left| M^B_{u,v} \right|$$

(7)

where $M^B_{u,v}$ represents the microblogs posted by user $u$ while $u$ performs user diffusion behavior $B$ on user $v$ by those microblogs. Note that we have extracted three types of user diffusion behaviors, thus $B$ can be $B_p$, $B_r$, or $B_m$.

After the attribute values of users are calculated, we normalize those values with (1).

**Reversed Behavior Activity:** The reversed behavior activity of a user expresses the relative activity of the user on receiving some kind of user diffusion behaviors performed by others, and its calculation is as follows:

$$A^B_{v,u} = \sum_{v \in V, v \neq u} \left| M^B_{v,u} \right|$$

(8)

After the attribute values of users are calculated, we normalize those values with (1).

**Retweeting Microblog Ratio:** When a user decides whether to retweet a microblog, the user has been exposed to the microblog via another user before. The retweeting microblog ratio between user $u$ and user $v$ reflects the fraction of the microblogs exposing to user $u$ via user $v$ are retweeted by user $u$, and can be calculated by the following formula:

$$O^B_{u,v} = \left| M^B_{u,v} \right| / \left| M^B_{v} \right|$$

(9)

where $M^B_{u,v}$ represents the microblogs posted by user $v$.

As described above, 11 properties belonging to three dimensions are captured to be used for predicting user retweeting behavior in Weibo. According to the above properties, various features can be generated from those identified properties of different heterogeneous diffusion relation networks, different kinds of users, and different types of user diffusion behaviors. We represent the set of these generated features as $F$.

Note that, we have identified three heterogeneous diffusion relation networks, three types of user diffusion behaviors, and three kinds of users (target user, parent user, and initial user) in this paper. The values of those various features are either integer values ($0$ or $1$) or numerical values (varying between $0$ and $1$).

B. PREDICTING USER RETWEETING BEHAVIOR ON MULTIPLE HETEROGENEOUS RELATION NETWORK

Combining all the above features generated from a multi-dimensional analysis of factors related to user retweeting behavior, we proposed a novel model – Retweeting Behavior on Multiple Heterogeneous Diffusion Relation Networks (RBMHDRN). RBMHDRN models the prediction of user retweeting behavior as an ensemble learning problem, i.e., incorporates the above features into a novel ensemble learning approach – Parallel Adaboost on Logistic Regression with a Single Feature (PALRSF) to predict user retweeting behavior.

Ensemble learning is one type of machine learning, using ensemble methodology to generate and combine multiple base models to solve a particular computational intelligence problem. It is well-known that ensemble methods can be used for improving machine learning performance [45], [46]. Many researches have employed ensemble learning methods to study problems.

The proposed novel ensemble learning approach PALRSF parallel integrates Adaboost [47], [48] with various Logistic Regressions, each of which uses only a single feature, inferring the probability of user retweeting behavior. Fig. 2 schematically depicts our proposed novel ensemble learning approach, PALRSF.

As can be seen from Fig. 2, PALRSF is an iterative process that raises different base models trained by the same training set, and then combines those base models to form a stronger final model. Meanwhile, each base model of PALRSF is constructed as Logistic Regression with a Single Feature (LRSF), selected from all the possible LRSF trained in parallel, each of which chooses a different single feature to use.

At each iteration, PALRSF first assigns a weight value to each training sample. For example, at iteration $k$, the weight values of the training samples are as follows:

$$D^{(k)} = (d_1^{(k)}, d_2^{(k)}, \ldots, d_N^{(k)})$$

(10)
where $N$ represents the number of the samples included in the training set. Then, PALRSF builds all the possible LRSF in parallel and preserves the best performance one. For example, Logistic Regression with $i$-th feature $F_i$ infers the probability of user $v$ retweeting microblog $m$ after one of his/her parent users $u$ has retweeted or initially posted $m$ as follows:

$$P_{w_i}(v|u, m) = P_{w_i}(x) = \frac{1}{1 + e^{-w_i x}}$$

where $x$ represents the feature vector of the sample, $x_i$ is the value of $i$-th feature $F_i$, and $w_i$ is the corresponding coefficient of $x_i$.

When the iterative process stops, PALRSF will obtain numbers of base models, for example, if $k$-th base model LRSF selects $i$-th feature $F_i$ to use, then the base model can be described as follows:

$$g_k(x) = P_{w^{(k)}_i}(x)$$

In addition, the final model combines those base models:

$$G(x) = \sum_{k=1}^{K} \alpha_k Z_{w_k} \cdot g_k(x)$$

where $K$ is the number of base models, $\alpha_k$ is the coefficient of $k$-th base model, and $Z_{w_k}$ is the normalization factor chosen so that $\alpha_k$ will be a distribution.

Finally, PALRSF has adaptively selected proper ones from various features to use for the prediction task, as each base model has compared all the features and selected out a specific feature of the best performance from them.

C. LEARNING ALGORITHM

As described in Section VI-A, there are $|F|$ different features prepared to be selected for use in our proposed model RBMHDRN. For PALRSF, each base model is selected from $|F|$ possible LRSF trained in parallel. Constructing a LRSF can be formulated as learning an optimal solution and we adopt the maximum likelihood estimation method to achieve it. For example, to estimate the coefficient value $w^{(k)}_j$ of $j$-th feature used by $j$-th possible LRSF when constructing $k$-th base model, the log likelihood function can be defined as

$$L(w^{(k)}_j) = \sum_{i=1}^{N} d^{(k)}_j \cdot \log([P_{w_j^{(k)}(X_i)}]^{y_i}[1 - P_{w_j^{(k)}(X_i)}]^{1-y_i})$$

where $X_i$ represents the vector of features of $i$-th training sample, and $y_i$ is 1 if the sample is positive, or else is 0.

Different with classic Logistic Regression (LR) that treats each training sample equally, PALRSF assigns different weights to training samples. Moreover, to avoid over fitting, we add a regularization term to LRSF. Then the maximum likelihood estimation equals to finding the minimization of the novel likelihood function:

$$L'(w^{(k)}_j) = -L(w^{(k)}_j) + \frac{\gamma}{2N} \cdot (w^{(k)}_j)^2$$

where $\gamma$ is the parameter of the regularization term.

To learn the coefficient value making the novel likelihood function minimized, the Gradient Descent method is adopted, while the factor of gradient is shown in the following formula:

$$\frac{\partial L'}{\partial w^{(k)}_j} = -\sum_{i=1}^{N} d^{(k)}_j (P_{w^{(k)}_j}(X_i) - y_i) X_{ij} + \frac{\gamma}{N} \cdot w_j$$

Thus, the coefficient value $w^{(k)}_j$ can be updated on the gradient direction with the learning rate $\beta$ iteratively.

$$w^{(k)}_j = w^{(k)}_j - \beta \cdot \frac{\partial L'}{\partial w^{(k)}_j}$$

Then, after $k$-th base model is obtained, the values of $\alpha_k$ and $D^{(k+1)}$ are updated as follows [47]:

$$\varepsilon_k = \sum_{i:j \neq j \neq j} d^{(k)}_j \frac{1}{2 \ln \frac{1 - \varepsilon_k}{\varepsilon_k}}$$

$$\alpha_k = \frac{1}{Z_D} \cdot \exp(-\alpha_k \gamma y g'_k(X_i))$$

where $\varepsilon_k$ is the weighted error of $k$-th base model, $Z_D$ is a normalization factor ensuring that $D^{(k+1)}$ will be a distribution, and $g'_k(X_i)$ is 1 when $g_k(X_i) > 0.5$, or else is $-1$.

According to Adaboost, generally the terminal condition of iterations is set as a specified number of iterations or an acceptable error range. However, this may result in over fitting. To avoid this problem, we further partition the dataset used to train and test our proposed model into three subsets, i.e., the training set, the test set, and the validation set, then use the training set to learn parameters and use the validation set to decide whether or not to terminate the iterations. At the end of each iteration, the error of the final model on the validation set can be obtained, and if the error does not decrease until next $Q$ iterations, the iterative training process is terminated and the ultimate final model is generated. In our experiments, for the final model, we set the iteration check interval $Q$ as 50.

According to [48], [49], our proposed approach PALRSF also can be interpreted as stagewise estimation procedures for minimizing an exponential objective function, and the error of the final model can drop exponentially fast. Therefore, our proposed approach PALRSF can converge to global optimal solution fast.

The pseudo-code of the algorithm for learning parameters of the proposed ensemble learning approach PALRSF is shown in Algorithm 1. The input of the algorithm is as follows. $X$ denotes the feature vectors of training samples, $X'$ denotes the feature vectors of validation samples, $Y$ is the labels of training samples (label is 1 for a positive sample, $-1$ for a negative sample), $Y'$ is the labels of validation samples, and $J$ represents the number of iterations for training a base model.
As the pseudo-code shows, the time complexity of our proposed approach PALRSF may be expressed as $O(KNJ)$, due to that the complexity is dominated by learning the coefficient values of $K$ base models, each of which updates the corresponding coefficient value with (16) whose time complexity is $O(NJ)$ times. Moreover, as we adopt Adam SGD [50] method instead of Gradient Descent method in practical experiments, we found that our approach could achieve high performance with $J < 3$ for learning coefficient values of each base model, thus the final time complexity of our proposed approach PALRSF can be reduced to $O(KN)$.

V. EXPERIMENTS

On a real world dataset, we compare our proposed prediction model RBMHDRN with two simple baselines, and one of the state-of-the-art models – LRC-BQ. In addition, we also compare our proposed model with other three state-of-the-art models, using the performance improvements over LRC-BQ as metrics. Then, we compare our proposed novel ensemble learning approach PALRSF with three classic methods, especially with two of them when using various sizes of the training set. The average of the performance metrics including precision, recall, F1-score, and accuracy on a 10-fold cross test are reported. Moreover, we evaluate the speedup performance of RBMHDRN as well.

A. DATASET DESCRIPTION

We used a dataset which contains 762,936 microblogs published by 68,817 users covering a period from September 9, 2009 to June 6, 2014 crawled from the famous Chinese microblog platform – Weibo (http://weibo.com), while about 99% of these microblogs were published after November 1, 2013. Because of the limitation set by the platform, the dataset only includes public microblogs that did not violate user privacy. The statistical properties of the dataset are shown in Table 2.

FIGURE 3. The complementary cumulative distribution (CCDF) of cascade depth.

| Users | Original Posts | Retweets | Direct Retweeting Relations | Indirect Retweeting Relations |
|-------|----------------|----------|----------------------------|------------------------------|
| 68,817 | 425,213        | 337,723  | 71,974                     | 400,469                     |

TABLE 2. Statistical properties of the dataset crawled from Weibo.
In this paper, each diffusion instance refers to a retweeting microblog between a parent user and a target user. Moreover, the diffusion instance can be used to generate a non-diffusion instance between the same parent user and another target user who has direct retweeting relationship to the parent user but has not retweeted the microblog in 30 days after the specified time window.

Using the instances generated from the later subset, we trained and tested the prediction model RBMHDRN with a 10-fold cross test, during which the instances were randomly partitioned into 10 subsets of approximately equal size. At each iteration, 8 subsets served as the training set including the samples used for training, 1 subset served as the validation set including the samples used for validation, and the remaining 1 subset served as the test set including the samples used for testing. This process was repeated 10 times, each time a different subset of instances was treated as the test set. We report the average of the evaluation metrics on those test sets as our final results.

Note that, in those practical experiments, we adopt Adam SGD [50] method instead of Gradient Descent method both in LR and PALRSF.

B. EXPERIMENTAL SETUP

To verify the effectiveness of our proposed prediction model RBMHDRN, we compare RBMHDRN with two simple baselines. One is based on the historical retweeting behaviors of the target user, i.e., whether the target user has retweeted another microblog initially posted by the same initial user in the past, called HRB. And the other is based on the empirical retweeting probability between the target user and the parent user, i.e., the fraction of microblogs exposing to the target user via the parent user are retweeted by the target user, called ERP.

We also compare our proposed model with one of the state-of-the-art models - LRC-BQ [29], with which several other state-of-the-art models are also compared by researches, such as GruBa [33] and M_SGM [28]. To conduct the experiment, we collected the necessary properties and implemented LRC-BQ to make it run on the dataset described in Section V-A. Furthermore, according the best performance reported in [33], [28], and [30], we compare our proposed model with other three state-of-the-art models – GruBa, M_SGM, and ASC-HDP [30], using the performance improvements over LRC-BQ as metrics.

Then, to examine the effectiveness of our proposed ensemble learning approach PALRSF, we compare PALRSF with 3 classic methods including Naive Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM), using the same features described in Section VI-A. Meanwhile, we plot the accuracies of PALRSF and LR during their training processes, thus to show the effectiveness of our integration of parallelized Adaboost with regularized Logistic Regression and making each base model use only one single feature for improving the performance of Logistic Regression.

### Table 3. Prediction performance of RBMHDRN and LRC-BQ.

| Model     | Precision | Recall | F1-score |
|-----------|-----------|--------|----------|
| RBMHDRN  | 92.96%    | 91.20% | 91.90%   |
| LRC-BQ    | 71.23%    | 88.13% | 78.78%   |

To report how the performance of each method is affected by the size of the training set, we also conducted experiments on the dataset with different training data settings. In these experiments, the validation set and the test set each contains 10% of the samples, while the training set can contain at most 80% of the samples. We report the performance of our proposed approach PALRSF, LR, and SVM using the same features described in Section IV-A when 10%, 30%, 50%, 70%, and 80% of the samples are used for training.

Moreover, we evaluate the speedup performance of RBMHDRN by increasing the number of threads from 2 to 8 with an increment of 2.

C. EXPERIMENT RESLUT AND ANALYSIS

As shown in Fig. 4, the F1-scores of two baselines - the historical retweeting behavior (HRB) and the empirical retweet probability (ERP) are roughly around 0.2, with relatively low recall scores. And ERP achieves the highest precision score, which might be due to that the empirical probability is naturally suitable for predicting occurrence probability, but still not suitable for predicting non-occurrence probability. Our proposed model RBMHDRN achieves the best performance, as its most measures are much better than that of those two baselines. The results demonstrate the effectiveness of our generated features and proposed ensemble learning approach.

Table 3 shows the performance of our proposed model RBMHDRN against one of the state-of-the-art models - LRC-BQ on the dataset described in Section V-A. As can be seen from the result, our proposed model outperforms the social influence locality based model regarding all the metrics. This indicates that our proposed model is benefiting from modeling user retweeting behaviors on multiple heterogeneous diffusion relation networks instead of a single network, using various features generated from the identified properties of different heterogeneous diffusion relation networks, different kinds of users and different types of user diffusion behaviors, and adopting a novel ensemble learning approach.

Several other state-of-the-art models have also been compared with LRC-BQ by researches on predicting user retweeting behaviors, such as ASC-HDP [30], GruBa [33], and M_SGM [28]. According the best performance reported in [30], we compare our proposed model RBMHDRN with ASC-HDP [30] in the performance improvements over LRC-BQ in terms of precision, recall, and F1-score. As can be seen from the result shown in Table 4, our proposed model can obtain high performance improvements over LRC-BQ in terms of precision and F1-score, which is comparable to
another state-of-the-art model ASC-HDP, but get a bit worse performance improvements over LRC-BQ than ASC-HDP in terms of recall. The possible reason is that our proposed model is constructed on multiple heterogeneous diffusion relation networks, and incorporates more features belonging to the dimensions of user diffusion behavior and network structure, but ignores the temporal information incorporated by ASC_HDP.

According to the results of performance comparisons with LRC-BQ reported in [33] and [28], GruBa can improve the accuracy over LRC-BQ by 6% [33], and M_SGM can achieve a 3.3-12.6% improvement compared with other six methods including LRC-BQ in terms of precision [28]. Meanwhile, our proposed model RBMHDRN, which is able to achieve 91.96% accuracy and 92.62% precision, can improve the accuracy over LRC-BQ by 15.69%, and achieve a 21.39% improvement compared with LRC-BQ in terms of precision. This strongly indicates that, our proposed model RBMHDRN can obtain better performance improvements over LRC-BQ than the above two state-of-the-art models Gruba and M_MSG in terms of accuracy and precision, respectively, for predicting user retweeting behavior in social networks. This may due to the fact that each of the above two state-of-the-art models ignores some important features. For M_MSG, some features of diffusion relation networks and user diffusion behaviors are ignored. For GruBa, some features of user mentioning behavior and indirect retweeting network are ignored. However, it is worth to note that both the emotional information considered by M_MSG, and the user clustering method and the basic features adopted by GruBa, can significantly improve the prediction performance, which may be important guiding to our future researches.

Therefore, it can be concluded that, our proposed model RBMHDRN is comparable to (sometimes better than) several state-of-the-art models in predicting user retweeting behavior in social networks.

We also present the best reported results of three state-of-the-art models in Table 5. However, this is only for inference due to the differences in training/testing data splits and data processing methods, which may make the direct comparisons of those results unfair. For example, ASC-HDP is reported to perform much better than LRC-BQ on the corresponding dataset and processing methods described in [30], while the best reported results shown in Table 5 could lead to the opposite conclusion.

As can be seen from Table 3 and 5, it is worth noting that the performance of LRC-BQ is also improved significantly compared with that in [29], by regarding the direct retweeting relationship between users instead of the following relationship between users as the main social relationship that supports the process of information diffusion in online social networks.

The comparison of our proposed ensemble learning approach and three classic methods is presented in Fig. 5, the results show that the performance of the prediction task based on Logistic Regression (LR) is better than that based on Naive Bayes (NB) using the same features described in Section IV-A, which achieves accuracy of 90.57% and 84.91% respectively. This may because that Logistic Regression is natural suitable for predicting diffusion probability. Support Vector Machine (SVM) achieves 90.66%
FIGURE 6. Plot of accuracies of PALRSF and LR.

FIGURE 7. Prediction performance of PALRSF and two classic methods with respect to the size of the training set.

The accuracy trend of LR and PALRSF can be seen from Fig. 6. We found that LR converges faster than PALRSF, this may due to the fact that PALRSF adopts weak LR as base model, which uses only one single feature, and PALRSF needs to generate numbers of weak LR, while LR uses all the features at once and needs to generate only one LR. However, PALRSF achieves 1.39% improvement over LR in terms of accuracy, which shows that integrating parallelized Adaboost with regularized Logistic Regression and making each base model use only one single feature, can effectively improve the performance of Logistic Regression.

Fig.7 presents the effect of the size of the training set on the prediction performance of our proposed approach PALRSF, SVM, and LR, using the same features described in Section IV-A. An upward trend can be seen in prediction accuracy as the size of the training set increases. With different sizes of the training set, our proposed approach PALRSF is able to achieve the highest score on accuracy. For SVM, the prediction performance is a bit relatively worse when 10% of the total instances are available for training. The reason for this may be that, the number of the training samples is not sufficient for SVM to extract steady pattern of user retweeting behavior. We can observe that, the accuracies of the above three methods increase only slightly when the size of the training set is increased beyond 50% of the total instances. This indicates that, once the size of the training set is sufficient to capture steady pattern for those methods to model user retweeting behavior, increasing the size of the training set further will only result in minor accuracy gain but much higher the cost of computing.

Furthermore, the accuracy of PALRSF starts to drop slightly when the size of the training set reaches 50% of the total instances. One possible reason is that, after that point, our proposed approach may overfit slightly due to the large number of the training samples. This may indicate that, our proposed approach PALRSF can achieve its best performance using less training samples than SVM and LR need, while significantly outperforms them. Moreover, the number of the samples used for training and validation is supposed to be set as 6 times the number of the test samples (1 for validation and 5 for training), which seems to be a reasonable choice of the size of the training set to achieve the best performance for predicting user retweeting behavior with PALRSF.

Moreover, the experimental results illustrated in Fig. 8 show that our proposed model RBMHDRN scales well with the number of threads, profiting from parallelization.

VI. CONCLUSION

In this paper, we have proposed a novel model based on multiple heterogeneous diffusion relation networks for the prediction of user retweeting behavior occurring in the process of information diffusion in social networks, by incorporating a novel ensemble learning approach. To infer the diffusion probability, we consider multiple heterogeneous diffusion relation networks while other researches usually consider
only one network. We identify a set of properties belonging to three dimensions — content semantics, user diffusion behavior, and network structure, to generate various features from those properties of different diffusion relation networks, different kinds of users, and different types of user diffusion behaviors.

Then, we treat our prediction problem as an ensemble learning problem and propose a novel ensemble learning approach PALRSF to solve the problem. Combing the generated features and our proposed novel ensemble learning approach, we propose a model RBMHDRN to predict user retweeting behavior in online social networks. Executed on a real world dataset of microblogs extracted from Weibo, the extensive experiments demonstrate the effectiveness of our generated features and proposed ensemble learning approach.

In the future, several works remain. The performance of the prediction problem still has space for improvement, considering broader and more detailed factors and introducing other efficient algorithms may be helpful. Various hypothesis tests can be performed on the training and testing samples, such as t-test, Kolmogorov–Smirnov test, and likelihood-ratio test, which are beneficial in terms of implying whether the values of a feature differ between positive samples and negative samples, or assessing the contribution of features to our proposed prediction model. Besides that, combining the other learning methods to explore the diffusion data in real time is also a challenge. In addition, studies like those above will give a richer understanding of how information diffuses in social networks.

REFERENCES

[1] M. Gupta, J. Gao, C. X. Zhai, and J. Han, “Predicting future popularity trend of events in microblogging platforms,” in Proc. 75th Ann. Meeting Amer. Soc. Inform. Sci. Technol. (ASIS&T), 2012, pp. 1–10.

[2] C. M. K. Cheung and D. R. Thadani, “The impact of electronic word-of-mouth communication: A literature analysis and integrative model,” Decis. Support Syst., vol. 54, no. 1, pp. 461–470, Dec. 2012.

[3] J. Borge-Holthoefer, R. A. Banos, S. Gonzalez-Bailon, and Y. Moreno, “Cascading behaviour in complex socio-technical networks,” J. Complex. Netw., vol. 1, no. 1, pp. 3–24, Jan. 2013.

[4] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida, “On word-of-mouth based discovery of the Web,” in Proc. ACM SIGCOMM Conf. Internet Meas. Conf. IMC, 2011, pp. 381–396.

[5] S. Pei, L. Muchnik, J. S. Andrade, Z. Zheng, and H. A. Makse, “Searching for superspreaders of information in real-world social media,” Sci. Rep., vol. 4, no. 1, pp. 5547–5558, May 2015.

[6] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, “The structural virality of online diffusion,” Manage. Sci., vol. 62, no. 1, pp. 180–196, 2015.

[7] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in Twitter,” in Proc. 20th Int. Conf. Companion World Wide Web WWW, 2011, pp. 57–58.

[8] H. W. Hethcote, “The mathematics of infectious diseases,” SIAM Rev., vol. 42, no. 4, pp. 599–653, 2000.

[9] X. X. Zhan, C. Liu, G. Zhou, Z.-K. Zhang, G.-Q. Sun, J. J. H. Zhu, and Z. Jin, “Coupling dynamics of epidemic spreading and information diffusion on complex networks,” Appl. Math. Comput., vol. 332, pp. 437–448, Sep. 2018.

[10] T. H. Kwan, N. S. Wong, G. C. Y. Liu, K. C. W. Chan, O. T. Y. Tsang, W. S. Leung, K. M. Ho, M. P. Lee, W. Lam, S. N. Chan, D. P. C. Chan, and S. S. Lee, “Incorporation of information diffusion model for enhancing analyses in HIV molecular surveillance,” Emerg. Microbes Infections, vol. 9, no. 1, pp. 256–262, Jan. 2020.

[11] M. Granovetter, “Threshold models of collective behavior,” Amer. J. Sociology, vol. 83, no. 6, pp. 1420–1443, May 1978.

[12] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, “Outtweeting the Twitterers—predicting information cascades in microblogs,” in Proc. 3rd Workshop Onl. Soc. Netw., Boston, MA, USA, Jun. 2010, pp. 10–11.

[13] I. Goldenberg, B. Libai, and E. Muller, “Talk of the network: A complex systems look at the underlying process of word-of-mouth,” Marketing Lett., vol. 12, no. 3, pp. 211–223, 2001.

[14] K. Saito, M. Kimura, K. Ohara, and H. Motoda, “Selecting information diffusion models over social networks for behavioral analysis,” in Proc. ECML-PKDD, Barcelona, Spain, 2010, pp. 180–195.

[15] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Discovery Data Mining KDD, 2003, pp. 137–146.

[16] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, “Thirty years of machine learning: The road to Pareto-optimal wireless networks,” IEEE Commun. Surveys Tuts., early access, Jan. 13, 2020, doi: 10.1109/ COMST.2020.2965856.

[17] J. Hu, Y. Luo, and J. Yu, “An empirical study on selectivity of retweeting behaviors under multiple exposures in social networks,” J. Comput. Sci., vol. 28, pp. 228–235, Sep. 2018.

[18] J. Shi, K. K. Lai, P. Hu, and G. Chen, “Understanding and predicting individual retweeting behavior: Receiver perspectives,” Appl. Soft Comput., vol. 60, pp. 845–857, Nov. 2017.

[19] X. Chen, X. Zhou, T. Sellis, and X. Li, “Social event detection with retweeting behavior correlation,” Expert Syst. Appl., vol. 114, pp. 516–523, Dec. 2018.

[20] Q. Li and Y. Liu, “Exploring the diversity of retweeting behavior patterns in chinese microblogging platform,” Inf. Process. Manage., vol. 53, no. 4, pp. 945–962, Jul. 2017.

[21] D. M. Romero, B. Meeder, and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter,” in Proc. 20th Int. Conf. World Wide Web WWW, 2011, pp. 695–704.

[22] G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” Commun. ACM, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[23] J. Wang, C. Jiang, T. Q. S. Quek, X. Wang, and Y. Ren, “The value strength aided information diffusion in socially-aware mobile networks,” IEEE Access, vol. 4, pp. 3907–3919, 2016.

[24] S. A. Myers and J. Leskovec, “Clash of the contagions: Cooperation and competition in information diffusion,” in Proc. IEEE 12th Int. Conf. Data Mining, Dec. 2012, pp. 539–548.

[25] L. Liu, X. Wang, Y. Zheng, W. Fang, S. Tang, and Z. Zheng, “Homogeneity trend on social networks changes evolutionary advantage in competitive information diffusion,” New. J. Phys., vol. 22, no. 1, pp. 13–19, Jan. 2020.

[26] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?” in Proc. 23rd Int. Conf. World Wide Web - WWW, 2014, pp. 925–936.

[27] D. Li, S. Zhang, X. Sun, H. Zhou, S. Li, and X. Li, “Modeling information diffusion over social networks for temporal dynamic prediction,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 9, pp. 1985–1997, Sep. 2017.

[28] J. Chen, Y. Liu, and M. Zou, “User emotion for modeling retweeting behaviors,” Neural Netw., vol. 96, pp. 11–21, Dec. 2017.

[29] J. Zhang, B. Liu, J. Tang, C. Chen, and J. Li, “Social influence locality for modeling retweeting behaviors,” in Proc. 23rd Int. Joint Conf. Artif. Intell. (IJCAI), Beijing, China, 2013, pp. 2761–2767.

[30] Q. Zhang, Y. Gong, Y. Guo, and X. Huang, “Retweet behavior prediction using hierarchical Dirichlet process,” in Proc. 29th AAAI Conf. Artif. Intell. (AAAI), Austin, TX, USA, 2015, pp. 403–409.

[31] C. Jiang, Y. Chen, and K. J. R. Liu, “Evolutionary dynamics of information diffusion over social networks,” IEEE Trans. Signal Process., vol. 62, no. 17, pp. 4573–4586, Sep. 2014.

[32] C. Jiang, Y. Chen, and K. J. R. Liu, “Graphical evolutionary game for information diffusion over social networks,” IEEE J. Sel. Topics Signal Process., vol. 8, no. 4, pp. 524–536, Aug. 2014.

[33] J. Zhu, S. Ma, H. Zhang, C. Hu, and X. Li, “Incorporating user grouping into retweeting behavior modeling,” in Proc. 23rd Int. Conf. Database Syst. Adv. Appl. (DASFAA), Gold Coast, QLD, Australia, 2018, pp. 479–490.

[34] E. Bakshy, I. Rosem, C. Marlow, and L. A. Adamic, “The role of social networks in information diffusion,” in Proc. 21th Int. Conf. World Wide Web (WWW), Lyon, France, 2012, pp. 519–528.
[35] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, “A survey of heterogeneous information network analysis,” IEEE Trans. Knowl. Data Eng., vol. 29, no. 1, pp. 17–37, Jan. 2017.

[36] C. Wang, R. Raina, D. Fong, D. Zhou, J. Han, and G. Badros, “Learning relevance from heterogeneous social network and its application in online targeting,” in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. SIGIR, 2011, pp. 655–664.

[37] J. Tang, W. Li, A. K. Usadi, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, and P. Xu, “PatentMiner: Topic-driven patent analysis and mining,” in Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD, 2012, pp. 1366–1374.

[38] X. Yin, J. Han, and P. S. Yu, “Object distinction: Distinguishing objects with identical names,” in Proc. IEEE 23rd Int. Conf. Data Eng., Apr. 2007, pp. 1242–1246.

[39] C. Wang, J. Han, Y. Jia, J. Tang, D. Zhang, Y. Yu, and J. Guo, “Mining advisor-advisee relationships from research publication networks,” in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD, 2010, pp. 203–212.

[40] C. Shi, Z. Zhang, P. Luo, P. S. Yu, Y. Yue, and B. Wu, “Semantic path based personalized recommendation on weighted heterogeneous information networks,” in Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. CIKM, 2015, pp. 453–462.

[41] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage., vol. 24, no. 5, pp. 513–523, Jan. 1988.

[42] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[43] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, “K-core organization of complex networks,” Phys. Rev. Lett., vol. 96, no. 4, pp. 229–252, Feb. 2006.

[44] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” Comput. Netw. ISDN Syst., vol. 30, nos. 1–7, pp. 107–117, Apr. 1998.

[45] T. G. Dietterich, “Ensemble methods in machine learning,” in Proc. 1st Int. Workshop Multi. Classifier Syst., Cagliari, Italy, 2000, pp. 1–15.

[46] L. Rokach, “Ensemble-based classifiers,” Artif. Intell. Rev., vol. 33, nos. 1–2, pp. 1–39, 2010.

[47] Y. Freund, R. E. Schapire, and N. Abe, “A short introduction to boosting,” J. Japn. Soc. Artif. Intell., vol. 14, no. 5, pp. 771–780, 1999.

[48] Y. Freund and R. E. Schapire, “A decision-theoretic generalization of online learning and an application to boosting,” J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, Aug. 1997.

[49] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: A statistical view of boosting,” Ann. Statist., vol. 28, no. 2, pp. 337–407, 2000.

[50] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, 2015, pp. 1–15.

LONG CHEN received the B.S. and M.S. degrees from the College of Computer Information and Engineering, Jiangxi Normal University, China. He is currently pursuing the Ph.D. degree with the South China University of Technology, China, under the supervision of Prof. H. Deng. His research interests include keyword extraction, information diffusion, and machine learning.

HUIFANG DENG received the B.Sc. and M.Sc. degrees in China and the Ph.D. degree from University College London (UCL), U.K. From 1989 to 2004, he was studying, working, and living in the U.K. (for nearly 16 years). From 2001 to 2004, he was a Chief Technical Officer and a Chief Scientist with Sunrise Systems Ltd., Cambridge, U.K. He was the Dean of the Software School, South China University of Technology, Guangzhou, China, in September 2004, where he is currently a Full Professor. He has published over 130 articles published and holds many patents for invention. His research interests include RFID technology and applications, the Internet of Things, cloud computing, big data, AI, machine learning, data mining, data fusion, neural networks, social computing, computer modeling, high-performance computing, and scientific computing.

* * *