 Retweeting Prediction Based on Social Hotspots and Dynamic Tensor Decomposition

Qian LI††, Xiaojuan LI†, Bin WU†, Nonmembers, and Yunpeng XIAO††, Member

SUMMARY In social networks, predicting user behavior under social hotspots can aid in understanding the development trend of a topic. In this paper, we propose a retweeting prediction method for social hotspots based on tensor decomposition, using user information, relationship and behavioral data. The method can be used to predict the behavior of users and analyze the evolution of topics. Firstly, we propose a tensor-based mechanism for mining user interaction, and then we propose that the tensor be used to solve the problem of inaccuracy that arises when interactively calculating intensity for sparse user interaction data. At the same time, we can analyze the influence of the following relationship on the interaction between users based on characteristics of the tensor in data space conversion and projection. Secondly, time decay function is introduced for the tensor to quantify further the evolution of user behavior in current social hotspots. That function can be fit to the behavior of a user dynamically, and can also solve the problem of interaction between users with time decay. Finally, we invoke time slices and discretization of the topic life cycle and construct a user retweeting prediction model based on logistic regression. In this way, we can both explore the temporal characteristics of user behavior in social hotspots and also solve the problem of uneven interaction behavior between users. Experiments show that the proposed method can improve the accuracy of user behavior prediction effectively and aid in understanding the development trend of a topic.

key words: social network, hotspot topic, behavior analysis, retweeting prediction, tensor decomposition

1. Introduction

With the development of Web 2.0, microblogs, blogs, forums, and other colorful social media [1] continue to emerge and mature. The emergence of these social media provides the possibility of sharing resources, exchanging information and makes the internet to maintain substantial amounts of user behavioral data. These data are rich and valuable, and mining useful information from massive data has become a popular research. We can acquire an understanding of the distribution [2] of public behavior based on analysis of user behavior in the network, and this can assist in detecting abnormal behavior and provide reasonable bases for the study of public opinion control and influence evaluation.

In social networks, connections between users are established on the basis of their relationship concerns, and users can share information and knowledge. In addition, they can discuss topics of common interest at any time. Retweeting is one important means through which users interact. If a user retweets information regarding hotspots, his or her fans can read the information and compare it to their own interests to decide whether to retweet the message themselves. Retweeting of a message by a user fans promotes the dissemination of the information and creates a cascading effect [3] through the “user-fans” relationship. Figure 1 shows the participation of users at a certain stage of one hotspot topic. As can be seen from Fig. 1, after the rise of a hotspot topic, some users will retweet blogs related to the hotspot, enabling followers of these users to read these blogs. And the followers will decide whether to continue to retweet these blogs combined with their own interests and the influence of friends. We can explore the development direction of hotspots through the study of user behavior regarding those hotspots, and this can assist us in discovering potential for focusing attention on individual hotspots, and can provide a basis for public opinion control and emergency warnings.

In recent years, scholars have conducted extensive research on the information retweeting behavior on online social networks. According to different focuses of prediction tasks, prediction can be categorized as information centered and user centered [4]. User-centered retweeting prediction has substantial advantages in that it can incorporate individual differences and personal interests, thereby opening new space for in-depth study of information dissemination mechanism. Although user-centered retweeting prediction

Manuscript received November 8, 2017.
Manuscript revised December 27, 2017.
Manuscript publicized January 30, 2018.
†The authors are with the Beijing Key Laboratory of Intelligence Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China.
††The authors are with the Chongqing Engineering Laboratory of Internet and Information Security, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
a) E-mail: liqian2017@bupt.edu.cn
DOI: 10.1587/transinf.2017EDP27894

Fig. 1 Schematic diagram of user engagement in a hotspot topic.
has achieved significant research results, there are still some challenges:

1. Inaccurate calculation. There can be a huge number of data in social networks, but useful information is extremely sparse, creating difficulties for the precise calculation of attributes. And these attributes are important driving factors that affect user behavior.

2. Network dynamics. Online social networks, especially microblogs, have both static (scale-free [5] and small world [6]) and dynamic properties. In general, nodes and edges of the network change over time, while traditional research has considered networks as static, resulting in what is known as the static problem of dynamic network [7].

3. Topic timeliness. Hotspots in social networks typically undergo evolution in a process of generation, development, and extinction. In hotspots evolution, participation in different stages of the topic is uneven. Regarding uneven data, how to predict user behavior dynamically and periodically has become the primary difficulty in retweeting prediction research.

At present, attribute information modeling is an effective analytic method for retweeting prediction, and user interaction is the key to improving prediction effectiveness. The advantages of tensor decomposition in feature extraction can assist in exploring interactions between users. We use the superiority of tensor decomposition in the sparse data processing and implicit relation mining, combined with time decay function [8], to mine interaction between users under hot topics. Due to the timeliness of the topic, we predict user behavior at different times that combines time slices and discretization. We use a real dataset, Tencent microblog, to verify the effectiveness of our method.

The contribution of our study can be summarized as follows:

1. We construct a 3-order tensor model of “hot user”- “alternative user”-“interactive behavior” based on tensor decomposition in data space and projection to solve the problem of data sparsity. The three indices of 3-order tensor are hot user, alternative user and interactive behavior, respectively. The model can be used to explore potential impacts among users, which can assist in analyzing the influence of friends on user behavior.

2. Time decay function is introduced to optimize the calculation of user interaction. Using time decay function to fit user behavior dynamically not only further quantifies the evolution of user behavior regarding current hotspots, but also solves the problem of interactions between users fluctuating over time.

3. We propose a prediction method that involves user retweeting behavior based on logistic regression with time slices and discretization to solve the problem of uneven user interaction regarding hotspots. The method not only dynamically predicts the participation of users regarding hotspots, but also assists in understanding the development trend of hotspots, so as to provide evidence for public opinion control and network navigation.

The remainder of this paper is organized as follows. Section 1 introduces the background of the problem and the current status of research on it. Section 2 discusses work related to our study. Section 3 formulates the problem. Section 4 describes the proposed method and related learning algorithms. Section 5 presents our experimental results using a real-world dataset. Section 6 concludes the paper.

2. Related Work

In social networks, information propagation primarily involves retweeting behavior. User behavior can be predicted to learn the interests and behavioral rules of users, and we can analyze the impact of user behavior on information dissemination. At present, user retweeting prediction is prediction based on user past behavior, user group influence or mixed features. Here, we discuss the research on user retweeting prediction and the application of tensor decomposition in recent years.

Prediction based on user past behavior makes use of user behavior to mine the potential interests of users assuming that a user interests will not change over a short time. The basic factors that affect and enable prediction of user behavior have been analyzed and studied based on the above assumptions [9], and collaborative filtering has also been used to measure the preferences of users and the interactions between users and messages. However, this method has some limitations in dealing with the data sparsity and cold start problems. Subsequent research combined collaborative filtering with rich features [10]–[12]. Keywords and themes of microblogs have been extracted [13], alleviating the cold start problem caused by data sparsity to some extent. But the above researches ignore the influence of network dynamics.

Prediction based on user group influence usually assumes that user behavior is influenced primarily by close friends. Zhang J, Tang J, et al. established the logistic regression and factor graph models [14], [15] based on a tree-based network, indicating that user behavior is affected by friends to some extent. Retweeting prediction based on social network structure involves the study of hotspots spread in the network structure based on user access and triples [16], [17]. In Refs. [18], [19] studied how user behavior was affected by friends, according to the diversity of network structure and the strength of social relations. These researches are based on individual blogs, they can not analyze the user behavior in a particular scene. However, in Ref. [20] built a public influence model based on the behavior of users to predict user behavior and mine the development trend of hotspots. In the context of the topic, the model showed that the influence of conformity plays an important role in changing user behavior.

Prediction based on mixed features transforms the
problem of behavior prediction into a binary classification. This method selects features that affect user behavior, and then selects the appropriate classifier to train the classification model. Xu [21] et al. divided the characteristics that affect user behavior into social context, content and publisher-based, and trained a variety of classifiers (Decision Tree, SVM and Logistic Regression). They compared the effectiveness of various features with the method of feature exclusion and showed that features of social relations are more important than others. In Refs. [22], [23] constructed a classification model based on machine learning to predict user behavior. In Ref. [24] used a classification algorithm to diagnose rumors in the information dissemination network and discovered laws of rumors. Although the classification method was the simplest and most intuitive, it was interpretively weak, and depended on the choice and combination of features. But when extracting features, these studies usually assume that the network structure is static, which may lead to the loss of some useful information.

From Ref. [25] we can see that matrix and tensor decomposition are important tools for feature extraction. Matrix and tensor decomposition are widely used for dealing with sparse data and dimensionality reduction. In Ref. [26] predicted user behavior using non-negative matrix decomposition, and limited the decomposition of the process using social relationship strength to improve the precision of retweeting prediction. In Ref. [27] proposed a matrix decomposition model based on message clustering for retweeting prediction. Tensor decomposition was applied to data mining in Refs. [28], [29], and some effective methods for dealing with sparse data in a large high-dimensional dataset were proposed. In Ref. [30] proposed an algorithm of social recommendation based on user trust and tensor decomposition, and used an incremental tensor recommendation algorithm to improve the efficiency of a training model. In Ref. [31] improved a matrix decomposition model based on collaborative filtering and proposed an algorithm of tensor decomposition based on context information, which can improve recommendation quality. In Ref. [32] used stochastic matrix theory to reveal the rich behavior of the real world. In Ref. [33] proposed a time link prediction method combining bipartite graph with matrix and tensor decomposition, and showed that the time link prediction performance of the bipartite graph and tensor decomposition was better. The above studies reflect the universality of tensor decomposition, and the application prospects of tensor decomposition will broaden in dealing with sparse data.

3. Problem Definition

3.1 Related Definitions

In general, a social network can be represented as a directed network \( G = (N, E) \), where \( N \) represents the set of nodes, and \( E \subset V \times V \) represents the set of directed edges. Edge \( e_{ij} = (n_i, n_j) \in E \) represents the relationship between the nodes \( n_i \) and \( n_j \), and we define node \( n_j \) as the follower of node \( n_i \). If \( e_{ij} = (n_i, n_j) \in E \) and \( e_{ji} = (n_j, n_i) \in E \) exist simultaneously, with node \( n_j \) as the follower of node \( n_i \), and node \( n_i \) as the follower of node \( n_j \), we define this relationship as friend. Messages published by node \( n_i \) are visible to their followers. For predicting the behavior of followers, the basic definitions are given in the context of hotspots:

**Definition 1.** Hot user \( G_t^H \) and network of hot user \( G_t^H = (U_t^H, E_t^H) \) where \( U_t^H \) is the set of users who participate in hotspots time period \( t \), \( G_t^H = (U_t^H, E_t^H) \) is hot users and hot network in time period of hotspots, and \( E_t^H \) is the edge set of hot users. i.e. In other words, this is the relationship network of hot user \( E_U \subset U \times U \).

**Definition 2.** Alternative user \( V_t^A \) and network of alternative user \( G_t^A = (V_t^A, E_t^A) \) where \( V_t^A \) is the set of fans of hot users in hotspots time period. Further, \( V_t^A \) is candidate defined as the follower of hot user in hotspots time period. \( G_t^A = (V_t^A, E_t^A) \) is alternative users and network in hotspots time period, and \( E_t^A \) is the edge of alternative users. In other words, this is the relationship network of alternative user \( E_V \subset V \times V \).

**Definition 3.** All users \( W_t \) and network of all users \( G_t^W = (U_t, E_t) \) where \( W_t \) is the collection of hot and alternative users in hotspots time period, \( G_t^W = (U_t, E_t) \) is all users and networks in hotspots time period, and \( E_t \) is the edge of hot users and alternative users. In other words, this is the relationship network of hot and alternative users \( E_U \times V_t \subset (U \cup V_t) \times (U \cup V_t) \).

3.2 Problem Formulation

To describe the scientific problem in this paper formally, firstly, we assume that \( G_t^U = ((U \cup V_t), E_t^U) \) is the network of hotspots in time period \( t \) and that \( B = \{(b, w_t, t) | t \in \psi, w_t \in (U \cup V_t) \} \) is the past behavior of all users. Secondly, we define alternative users \( V_t^A \) and alternative network \( G_t^A = (V_t^A, E_t^A) \) from the complete network \( G_t^W = ((U \cup V_t), E_t^U) \). Finally, we predict \( R_t^{U+1} = (r_1, r_2, \ldots, r_t, \ldots) \), where \( r_t \) indicates whether alternative user \( v_t \) will participate in the next hotspots period. In other words, the problem is clearly defined as:

\[
G_{t+1}^U \rightarrow G_t^V \quad \Rightarrow f : (G_t^V) \rightarrow R_t^{U+1}
\]

3.2.1 Problem Input

Based on the relevant definitions in Sect. 3.1, the input of this paper is:

1. Relationship network of hot and alternative users \( G_t^W = ((U \cup V_t), E_t) \) in hotspots time period \( t \).
2. The past behavior information of all users \( B = \{(b, w_t, t) | t \in \psi, w_t \in (U \cup V_t) \} \), where \( (b, w_t, t) \) indicates that user \( w_t \) has published or retweeted information \( b \) in time period \( t \), and \( \psi \) is the time set of the latest month before the hotspots were launched.
3. The basic information of all users $D = \{(d, w_i) | w_i \in (U \cup V)\}$, where $(d, w_i)$ is the basic information of user $w_i$, such as the number of followers and friends.

3.2.2 Problem Output

Based on the above description, the problems to be solved are as follows:

1. How to construct a model to quantify the importance of different attributes for user retweeting behavior? We introduce a set of parameters $\theta$ to indicate the importance of different attributes. We determine the set of optimal parameters $\theta^* = \arg \max_\theta P_\theta(R^t|G, B)$, according to the user historical behavioral data.

2. How to predict individual behavior using the set $\theta^*$? The model with optimization parameters is leveraged to predict the behaviors of alternative users in hotspots time period $t+1$, that is, $P^* = \arg \max_{P^{t+1}} P_\theta(R^{t+1}|G, B)$.

4. Proposed Method

To solve the above problems, we propose a method of user retweeting prediction based on user information, behavioral and relational data. The details of this method are introduced in three stages, influence quantification, dynamic modeling, and retweeting prediction, as shown in Fig. 2. In the first stage, three factors affecting user behavior are quantified, and two driving features are defined to represent them.

In the second stage, we move to dynamic modeling based on the relationship data to analyze the influence of friends on user interaction. In the third stage, the time slices and discretization are used to deal with the hotspots life cycle, and the user retweeting prediction model is constructed to predict the behavior of the user and analyze the development trend of hotspots.

4.1 Influence Quantification

In social networks, a user will forward the content of hotspots or participate in a discussion of hotspots based on the influence of a friend. Sometimes, according to user interests, users will participate in a discussion of hotspots spontaneously. This shows that diverse factors affect user behavior, as shown in Fig. 3. These factors can be divided roughly into two categories, internal and external factors. We can determine the impact of different factors and find the key factors affecting user behavior through the quantification of different factors. Moreover, we also define two types of driving features to represent the classes of influential factors. The details are shown as follows:

4.1.1 Internal Driving Features

Whether a user would retweet a topic is usually related to the user own characteristics, such as user activity or interest. In this paper, the factors related to the user characteristics are defined as factors influenced by internal driving features. The specific description is given as follows:
1. user activity

Whether a user would retweet a topic is related to his or her activity on the social platform. In general, higher activity results in a greater possibility of the user participating in the current topic. The activity of alternative user $v_j$ is defined as:

$$activity(v_j) = \alpha \times origNum(v_j) + \beta \times retwNum(v_j)$$

where $origNum(v_j)$ and $retwNum(v_j)$ are the original and the retweeting numbers, respectively, of alternative users of the latest month before the topic was launched respectively. $\alpha, \beta$ are adjustable parameters with $\alpha, \beta \in [0, 1]$.

2. user individual factors

Whether a user would retweet information on a topic is related to his or her inherent attributes (number of fans or friends). We refer to these inherent attributes as user individual factors. In general, a greater number of fans or friends results in a greater possibility of the user participating in the current topic. The inherent attributes of alternative user $v_j$ include $Count of Fans(v_j)$: the number of fans of alternative user $v_j$, $Count of Idol(v_j)$: the number of friends of alternative user $v_j$, and $Count of HU(v_j)$: the number of participants users followed by alternative user $v_j$. Where, $Count of Fans(v_j)$ and $Count of Idol(v_j)$ can be obtained directly in the process of data crawling. And $Count of HU(v_j)$ can be obtained by counting the number of coincidence of $v_j$’s friends and hot users.

4.1.2 External Driving Features

Since a user participates in the discussion of a topic based on the influence of friends, and the influence of friends can be quantified by the interaction between users, we refer to the factors that affect user behavior as user interaction. As user interaction is quantified based on user relationships, it is defined as the factors influenced by external driving features. If the friend of a user has become a hot user, then a greater intensity of interaction between them results in a greater possibility of the user participating in the current topic. The interaction between the hot user and the alternative user is defined as:

$$strengthInteract(u_i, v_j) = I_{ij} \sum_{k=1}^{K} \sum_{h=1}^{3} interact(blog_{kh})$$

where $I_{ij} = \begin{cases} 1 & u_i \text{ is friend of } v_j \\ 0 & \text{others} \end{cases}$ is an indicator function, $blog_{kh}$ indicates that alternative user $v_j$ has retweeted or commented his friend’s ($u_i$) $k^{th}$ microblog, $K$ is the total number of microblogs of hot user $u_i$ and $interact(blog_{kh}) = \begin{cases} 1 & \text{alternative user } v_j \text{ participated in the hot user } u_i \text{'s } k^{th} \text{ microblog based behavior } b \\ 0 & \text{others} \end{cases}$.

4.2 Dynamic Modeling

In social networks, users interact with each other through social relations, such as retweeting and commenting, and users can share interests and information based on the social connection. However, the intensity of interaction between users will change with the growth of users life experience or changes in interests. That is, interaction between users in social networks has aging characteristics. These characteristics result in the interaction between users having the form of an exponential decline [8]. To quantify the influence of interaction on user behavior dynamically, an exponential decay function is introduced to optimize the calculation of user interaction:

$$strengthInteract(u_i, v_j) = I_{ij} \sum_{k=1}^{K} \sum_{h=1}^{3} interact(blog_{kh})$$

$$= I_{ij} \sum_{k=1}^{K} \sum_{h=1}^{3} I_{kh} e^{-(1+\kappa(\tau-t_k))}$$

where $I_{kh} = \begin{cases} 1 & \text{alternative user participate in hot user's } k^{th} \text{ microblog based behavior } b \text{ at } t \\ 0 & \text{others} \end{cases}$ is the current time of the hotspots, $t_k$ is the time that the hot user published the $k^{th}$ microblog, and $\kappa$ is the adjustable parameter.

The interactions between users is calculated based on explicit relationships, but the data sparsity problem of user interaction behavior is not solved. We construct the 3-order tensor $A \in R^{I\times J\times K}$ of “hot user-alternative user-interactive behavior”, and we take advantage of the characteristics of tensor decomposition in data space and projection to resolve the data sparsity of user interaction behavior, where $I$ is the dimension of hot users, $J$ is the dimension of alternative users, and $K$ is the dimension of interaction between users.
The tensor is shown in Fig. 4:

At the same time, we can mine implicit interactions between users with a tensor model based on user similarity, and the model can be used to analyze the impact of the relationship between friends on user behavior. In this paper, the high-order singular value decomposition (HOSVD) [34] model is used to factorize the user interaction tensor shown in Fig. 5. The 3-order user interaction tensor $A$ is decomposed into three factor matrices $U(1)^{r}, U(2)^{r}, U(3)^{r}$, and a core tensor $S$, as shown in Fig. 5, and the following calculations are required:

1. We apply HOSVD to a 3-order user interaction tensor $A$, and we define the three matrix unfolding operations as follows: $\mathbf{A}(1) \in \mathbb{R}^{k \times (jk)}, \mathbf{A}(2) \in \mathbb{R}^{k \times (jk)}, \mathbf{A}(3) \in \mathbb{R}^{k \times (ij)}$, where $\mathbf{A}(1), \mathbf{A}(2)$ and $\mathbf{A}(3)$ are called the 1-mode, 2-mode, and 3-mode matrix unfolding of $A$, respectively.

2. We apply singular value decomposition (SVD) to each $\mathbf{A}(n), 1 \leq n \leq 3$. The SVD decomposition of the 1-mode matrix unfolding is as follows:

$$
\mathbf{A}(1) = U_{1} \cdot \Sigma_{1} \cdot V_{K}^{T}
$$

The characteristic matrix $\Sigma$ is a diagonal matrix arranged according to the singular values of matrix $\mathbf{A}(1)$, where a greater singular value indicates greater importance of the feature to the matrix $\mathbf{A}(1)$. The matrix $\mathbf{A}(1)$ is usually described by the $c_{1}$ ($c_{1} < \min(J, K)$) most important features. After $c_{1}$ is determined, we use the left singular matrix $U$ and the right singular matrix $V$ to mine the latent information under implicit attention, resulting in the reconstruction matrix $\hat{\mathbf{A}}(1)$:

$$
\hat{\mathbf{A}}(1) = U_{1} \cdot \Sigma_{1} \cdot V_{K}^{T}
$$

Similarly, we can obtain

$$
\hat{\mathbf{A}}(2) = U_{2} \cdot \Sigma_{2} \cdot V_{K}^{T}
$$

$$
\hat{\mathbf{A}}(3) = U_{3} \cdot \Sigma_{3} \cdot V_{K}^{T}
$$

where $c_{2} (c_{2} < \min(J, K))$ and $c_{3} (c_{3} < \min(K, I))$ are the numbers of important features of matrix $A(2)$ and $A(3)$, respectively.

3. We calculate the core tensor based on the left singular matrix obtained in step 2.

$$
\mathbf{S} = \mathbf{A} \times_{1} U(1)^{T} \times_{2} U(2)^{T} \times_{3} U(3)^{T}
$$

4. We reconstruct the user interaction tensor based on the core tensor obtained in step 3.

$$
\hat{\mathbf{A}} = \mathbf{S} \times_{1} U(1)^{T} \times_{2} U(2)^{T} \times_{3} U(3)^{T}
$$

The element $(u, v, b, strength)$ of $\hat{\mathbf{A}}$ indicates that based on the behavior $b$, the interaction between hot user $u$ and alternative user $v$ is strength. We can calculate the interaction of hot and alternative users using the results of approximate tensor. The $N$ in Eq. (10) is the dimension of hot users.

$$
friendInteract(u, v_{j}) = \sum_{i=1}^{N} strengthInteract(u, v_{j})
$$

4.3 Retweeting Prediction

In addition to the influencing factors of the different driving features mentioned above, we propose a user retweeting prediction model based on logistic regression. The function is defined as:

$$
P(r|x) = \frac{1}{1 + e^{-\theta^{T} x}} = \frac{1}{1 + e^{-\theta_{0} + \theta_{1} x_{1} + \theta_{2} x_{2} + \theta_{3} x_{3} + \theta_{4} x_{4} + \theta_{5} x_{5}}}
$$

where $x_{1}$ refers to user activity, $x_{2}, x_{3}, x_{4}$ refers to user individual factors, and $x_{5}$ refers to user interaction. The parameters $\theta_{0}, \theta_{1}, \theta_{2}, \theta_{3}, \theta_{4}, \theta_{5}$ can be updated using gradient descent, through the parameter updating process

$$
\theta_{j-new} = \theta_{j-old} + \tau \left( r - \sum_{j=1}^{5} \theta_{j} x_{j} \right) \ast x_{j}
$$

where $\tau$ is the learning step, $r$ indicates whether the alternative user retweets the topic information. Regarding whether the parameters $\theta$ converge, convergence is considered to have occurred when the cost function is less than a threshold or the number of iterations reaches its maximum. If the parameters converge, we output the parameters. If they do not converge, then we continue to update the parameters until they do converge, in order to obtain the optimal parameter set $\theta^{*}$. Finally, retweeting behavior is predicted using the definition of logistic regression. When the value of $P(r|x)$ is greater than the threshold $\epsilon$, set $r = 1$; otherwise, set $r = 0$. 
The definition of $P(r|x)$ is:

$$r = \begin{cases} 
1 & P(r|x) \geq \varepsilon \\
0 & \text{others} 
\end{cases}$$

(13)

It is assumed that when $r = 1$, the alternative user will retweet the hotspots information in the next period; otherwise, the alternative user will not retweet the topic information.

4.4 Learning Algorithm

Complete networks of hotspots and past behaviors are provided, and the attribute parameters $\alpha, \beta$, learning step $\tau$, and classification thresholds $\varepsilon$ are initialized. Our goal is to calculate the optimal parameters $\theta^*$ and the results of retweeting prediction $P(r|x)$. The learning algorithm is as follows:

Learning parameters are initialized for the learning algorithm, and then user activity, individual factors and user interaction are calculated, respectively. Finally, the user behavior is predicted using logistic regression. Time decay function is introduced to calculate the interactions between users, and then, we construct the "hot user"-"alternative user"-"interactive behavior" tensor model to mine the implicit interactions between users, where, the time complexity of calculating user activity is $O(J)$. And the time complexity of calculating user individual factors is $O(I \ast J)$. When user interaction is calculated based on tensor decomposition, the resource cost is concentrated primarily in the process of decomposition, and the complexity is $O(R \ast I \ast J \ast K) + O(S \ast I \ast J \ast K) + O(T \ast I \ast J \ast K)$, which is related to the size of the tensor. The prediction stage uses gradient descent update parameters, and time complexity is $O(J)$, resulting in an overall time complexity for the algorithm of $O(R \ast I \ast J \ast K) + O(S \ast I \ast J \ast K) + O(T \ast I \ast J \ast K) + 2O(J)$.

5. Experiments and Analysis

5.1 Experimental Settings

In this section, the experimental data are introduced first. Then, the baseline methods used in the experiments are presented. Finally, evaluation metrics are proposed to evaluate the performance of our method.

5.1.1 Experimental Data

The datasets in this paper are collected from Tencent microblog. In China, Tencent microblog is a very popular social network platform. It provides a possibility for people to share resources and interact with each other. To evaluate the performance of the hotspots behavior prediction model that our method used, we collected hotspots data from Tencent microblog, and we demonstrate the model on three different hotspots, Personal Tailor as topic A; Dad, Where Are We Going Season2 as topic B; Rare Blood Type as topic C. The total size of the experimental data is approximately 25G. Table 1 shows the statistics of the three topics.

Figure 6 describes the rate of original and retweeting of alternative users. It indicates that the behavior of retweeting is also one of the main means by which users participate in microblogging, and it also shows that the research on user forwarding behavior has some practical significance.

5.1.2 Baseline Methods

To evaluate the performance of the proposed method, this paper uses the following baseline methods for comparison with our method:

- Hawkes Process (HP) [19]: This method is a special linear self-excited method that can be used to extract user attribute information. We can use HP to study the relationship between users and the network structure, and it can be combined with logistic regression to predict user behavior.

| Dataset | Topic A | Topic B | Topic C |
|---------|---------|---------|---------|
| start time | 2013.12.19 | 2014.05.14 | 2014.02.25 |
| end time | 2014.01.04 | 2014.09.04 | 2014.09.07 |
| hot users | 3,504 | 7,022 | 9,626 |
| alternative users | 847,500 | 879,780 | 534,459 |
| edges | 1,365,106 | 1,075,703 | 582,815 |
| hot users behavior | 788,301 | 2,432,377 | 634,768 |
| alternative users behavior | 1,125,228 | 1,075,703 | 582,815 |

Algorithm 1: Learning algorithm

Input:
Users relationship network in time period $r$: $G^r_{IU} = ((U \cup V)^r, E^r_{IU})$, user past behavior: $B = \{(b, w_i) | b \in \psi, w_i \in (U \cup V)\}$, user information: $D = \{(d, w_i) | w_i \in (U \cup V)\}$.

Output:
Parameters: $\theta^* = \arg\max P_r(R^G(G, B))$.

// Initialization:
1. initialize $\alpha, \beta, \tau$ and $\varepsilon$;
2. calculate activity($v_i$) from Eq. (1);
3. calculate individual factors individual($v_i$);
4. calculate friendInteract($u, v_i$):
   - construct tensor;
   - tensor decomposition:
     - matrix unfolding $A_{1(1)}, A_{2(1)}, A_{3(1)}$ from Eqs. (5)-(7);
     - core tensor $S$ from Eq. (8);
     - reconstruct tensor $A$ from Eq. (9);
   - calculate friendInteract($u, v_i$) from Eq. (10);
5. calculate hypothesis function $P(r|x)$ from Eq. (11);
repeat
- update parameters $\theta$ from Eq. (12);
until convergence;
7. get parameter $\theta^* = \arg\max P_r(R^G(G, B))$;
8. predict $P^* = \arg\max_{P^*} P_r(R^{+1}(G, B))$ from Eq. (13).
Learning to Rank (LTR) [22]: This method can translate a scheduling problem into a classification problem in machine learning. The value obtained by judging the correlation between users and then using the scoring function can be used to predict user behavior.

Collaborative Filtering (CF) [10], [12]: Collaborative filtering is a model based on past behavior. This method assumes that user interests will not change for a short time, and user retweeting behavior is driven by their interests. It uses similarity of users to predict retweeting behavior.

Factor Graph (FG) [15], [16]: Factor Graph represents dependencies between functions and variables in the form of a bipartite graph. Each factor graph contains variable and factor nodes. It decomposes a probability function into various factors, and then deduce the probability based on the relationship of each factor. Depending on the probability obtained, user retweeting behavior can be predicted.

5.1.3 Evaluation Metrics

In this paper, we use Precision (Prec.), Recall (Rec.), F1-measure (F1), and receiver operating characteristic (ROC) curves to evaluate experimental results. In our case, if alternative users in time period $t$ are involved in the hotspots in time period $t+1$, then he or she is recorded as a positive sample (“1”), that is, he or she will retweet the microblog on this hotspots, and otherwise as a negative sample (“0”), that is, he or she will not retweet the microblog on this hotspots. The prediction results are expressed in the form of a confusion matrix, as shown in Table 2.

Table 2 Confusion matrix of prediction.

| actual class | predicted class |
|--------------|-----------------|
| involved (“1”) | TP (True “1”) | FN (False “0”) |
| not involved (“0”) | FP (False “1”) | TN (True “0”) |

Following evaluation criteria are defined: Precision measures the accuracy of the user behavior prediction model, defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$ (14)

Recall measures the comprehensiveness of the user behavior prediction model, defined as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$ (15)

F1 is a comprehensive evaluation index of precision and recall, defined as follows:

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$ (16)

An ROC curve can display overall precision and recall on the test sample intuitively, defined as follows:

$$\begin{align*}
\text{TPR} &= \frac{TP}{TP + FN} \\
\text{FPR} &= \frac{FP}{TN + FP}
\end{align*}$$ (17)

In this paper, the ROC is used to evaluate the impact of various influencing factors on the predicted result. An ROC curve takes the false (FPR) and true positive rates (TPR) as the abscissa and ordinate, respectively. The closer the ROC curve is to the upper left, the better the predicted result.

5.2 Performance Analysis

In this section, Observational Analysis is introduced first. Then, we carry out classifier selection and the analysis of the influencing factors. Finally, we evaluate the performance of the proposed method by comparing it with other baseline methods. And we fit the development trend of the topic.

5.2.1 Observational Analysis

Before conducting the comparative experiment, we verify that there is a practical significance to predicting the behavior of alternative users based on simple statistics. Figure 7 shows the distribution of participation in different hotspots, where the abscissa and the ordinate represent the time (in hours) and the number of participants, respectively. The original value is the participation of hot users in a particular hotspots period, and the fans join value is the participation of alternative users in a particular hotspots. From Fig. 7, it is easy to observe that the participation curves of alternative users are similar to the real curves in Topics A, B, and C. Thus, it is of practical significance to predict the development trend of the hotspots through predicting the behavior of
alternative users. At the same time, Fig. 7 shows that when the time period is 16-32 hours, the curve of the participation of alternative users is highly similar to the curve of hot users on topics A and B. When the time is 8-24 hours, the participation curves of alternative users are highly similar to the curves of hot users on topic C. Hence, the comparative experiment of this paper focuses primarily on the above period (the active period) of the hotspots.

5.2.2 Classifier Selection

Attribute information modeling is an effective analytic method for retweeting prediction. In order to maximize the role of property, it is crucial to select optimal parameters of attributes. We calculate user activity by changing the adjustable parameters, and we also predict the retweeting behavior of alternative users through user activity. In addition, some of the evaluation indicators are used to assess the prediction results. The optimal combination of parameters is selected based on observing the prediction results (taking topic A for example). Figure 8 compares the evaluation indicators of different activity parameters, where the abscissa indicates the adjustable parameters of the number of original microblogs or the number of retweeting microblogs, and the ordinate indicates the value of the evaluation metrics. It can be seen from Fig. 8 that when $\alpha = 0.5, \beta = 0.5$, the model that we used can achieve the best prediction effect. Hence, $\alpha = 0.5$ and $\beta = 0.5$ are used as the parameter values for calculating user activity.

We choose different classifiers to carry out experiments for selecting the appropriate classifier as our prediction model. As can be seen from Tables 3, 4 and 5, the results of each classifier are very similar. However, the effect of logistic regression is the best, so we choose logistic regression as the final prediction model.

5.2.3 Factor Contribution Analysis

We can determine the influencing factors of different driving features on the retweeting prediction model by reducing a certain factor. Figures 9, 10, and 11 show the results of different factors in the active period of the hotspots, where the abscissa and the ordinate indicate the time slices of the hotspots and the number of participants, respectively. When combining internal and external driving features to predict user behavior, as can be seen from Fig. 9 to Fig. 11, the predicted result is very close to the actual situation. However, when the single factor under a mechanism is used, there are some differences in the experimental results, and the results are less effective than those predicted by using all factors. According to the subgraph (a), it can be found that in the case of topic A, when user individual factors are used to predict user retweeting behavior separately, the prediction curve is closer to the real value, and user interaction is the second. Comparatively speaking, the prediction effect based on user activity is somewhat poor. By the observation and analysis of subgraph (b), we can find that the prediction effect of individual factors is better than that of user interaction, and the prediction effect of user interaction is superior to user activity. It can be found that topic B has some con-
Fig. 9 Comparison of prediction result and real value of all influencing factors and user activity.

Fig. 10 Comparison of prediction result and real value of all influencing factors and individual factors.

Fig. 11 Comparison of prediction result and real value of all influencing factors and user interaction.

Fig. 12 Comparison of evaluation indices of different influencing factors.

To evaluate the prediction effects of different influencing factors, this paper uses different evaluation indices and ROC curves to compare them. The comparison results are shown in Figs. 12 and 13. In Fig. 12, the abscissa and the ordinate indicate the different evaluation indices and the values of the evaluation indices, respectively. When using a single factor to predict user retweeting behavior, considering different evaluation metrics, user individual factors have the best prediction effect for topic A and Topic B. However, for topic C, user interaction is the best. In Fig. 13, the abscissa and the ordinate indicate the FPR and TPR, respectively. For topic A and topic B, using a single factor, the ROC curve based user individual factors is closest to the top.
For topic C, the ROC curve based user interaction is closest to the top left. According to the characteristics of Roc, we can make a conclusion that user individual factors play a leading role for topic A and Topic B, while for Topic C, user interaction plays a leading role. It can be found that the same conclusions are drawn from Fig. 9 to Fig. 13. At the same time. we can see from the figure that the prediction effect of using all influencing factors is better than that of using a single factor. And we can conclude that user activity does play some roles in actual prediction.

5.2.4 Predict Performance Analysis

The performance of the model that our method uses is evaluated by comparing the experimental results of the baseline methods. We use different metrics to evaluate the prediction effect of different algorithms. The results of the three different hotspots are shown in Tables 6, 7 and 8. As can be seen from the table, considering different evaluation metrics, the method we proposed has the best predict effect.

We predict the development trend of the active period of the hotspots by using the results of retweeting prediction. Figure 14 shows a comparison of the predicted results with the real results, where the abscissa and the ordinate represent the time slices of the hotspots cycle and the number of participants in different time slices of the hotspots, respectively. It can be seen from Fig. 14 that the model that our method used can assist effectively in enhancing understanding of the development trend of hotspots.

5.3 Discussion

This paper focus on modeling of retweeting prediction based on social hotspots. There are several important related issues that are discussed as follows.

1. observational analysis of influencing factors

Social networks may influence an individuals behavior, but they also reflect the individuals own activities and interests. In the context of our study, there are three possible driving factors that affect the behavior of the users. The driving factors refer to user activity, individual factors and interaction, respectively. In Sect. 5.2.3, the contribution of three driving factors is analyzed. Our experiment shows that the performance of a single influencing factor is different for hotspots mentioned in this paper. Also, we can find that the predic-
tion effect of using all influencing factors is better than that of using a single factor. In a follow-up study, it can be considered to label different hotspots, such as the label of entertainment, sports and others. Then, a general conclusion can be got for each type of hotspot.

2. comparison of prediction effect

In social network, current retweeting prediction can be categorized as user centered and information centered. This paper, from the view point of users, extracts the features that affect user behavior and constructs a logistic regression model to predict the behavior of alternative user in the next period of the hotspot. In Sect. 5.2.4, compare with the classical retweeting prediction model, the model this paper proposed has the best predict effect. However, the features related to information (such as the content, the languages of microblogs) has not been considered. Jenders [35] et al. use Natural Language Processing technology to analyze the content of microblog and find that content features and user characteristics can be complementary. When content and user features are used together, the prediction accuracy can be improved. In the future, we can further verify the validity of the model proposed in this paper by incorporating content features.

Until now, scholars have presented extensive research on retweeting prediction. There are still some questions to be explored in depth. For example, information overload causes the problem of information competition. And big data processing problems caused by massive data on social networks. In a word, retweeting prediction is a field full of challenges and opportunities.

6. Conclusion

Users were divided in this paper into hot and alternative users based on interactive data of social networks. We were able to predict the behavior of alternative users based on basic information, relational data, and previous behavioral data. Hot and alternative users were identified on the basis of a user time participating in a topic and the relationships between users. Then, time decay function was introduced to calculate the interaction strength between users using interaction data. The effects of user interaction were mined by using a model of tensor decomposition. Finally, we were able to predict the behavior of alternative users based on logistic regression combined with the factors mentioned.

This paper used the data of the social network (Tencent microblog) to conduct experiments. Experimental results demonstrated that the proposed method can improve the prediction accuracy rate, recall rate, and F1, and it outperformed baseline methods in terms of precision and recall measurement. In this paper, from the view point of users, driving factors that affect user behavior was analyzed. However, Text information such as the content and the languages of microblogs was ignored. There are two reasons why text information did not be used. On the one hand, text modeling is an important research content in natural language processing filed. On the other hand, the novelty of this paper would be in external feature extraction. The method proposed in this paper can assist more effectively in acquiring understanding of the development trend of a topic to provide evidence for public opinion control and emergency warnings.

Acknowledgements

This work is supported by the National Basic Research Program of China (Grant No. 2013CB329606), the National Natural Science Foundation of China (Grant No. 61772098), Chongqing Science and Technology Commission Project (Grant No. cstc2017jcjyAX0099) and the Foundation of Ministry of Education of China and China Mobile (Grant No. MCM20130351).

References

[1] K.L. Derby, “Social media: multiple channels to capture multiple audiences,” ACM SIGmoms Conference on User Services, pp.159–162, 2013.
[2] Y. Altshuler, E. Shmueli, G. Zyskind, O. Lederman, N. Oliver, and A. Pentland, “Campaign optimization through behavioral modeling and mobile network analysis,” IEEE Trans. Comput. Social Syst., vol.1, no.2, pp.121–134, 2015.
[3] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?,” International Conference on World Wide Web, pp.591–600, 2010.
[4] L. Li, Yang, J.H. Chen, and Z. Li, “Survey on predicting information propagation in microblogs,” Journal of Software, vol.27, no.2, pp.247–263, 2016.
[5] A. Barabasi and R. Albert, “Emergence of scaling in random networks,” Science, vol.286, no.5439, pp.509–512, 1999.
[6] W. DJ and S. SH, “Collective dynamics of small-world networks,” Nature, vol.393, no.6684, pp.440–442, 1998.
[7] C. Li, B.Q. Feng, Y.M. Li, S.L. Hu, N. Yang, and C.J. Tang, “Role-based structural evolution and prediction in dynamic networks,” Journal of Software, vol.28, no.3, pp.663–675, 2017.
[8] Z.M. Zhong, Y. Hu, C.H. Li, and Z.T. Liu, “Discovering similar users for specific user on microblog,” Chinese Journal of Computers, vol.39, no.4, pp.765–779, 2016.
[9] B. Jiang, J.G. Liang, Y. Sha, et al., “Retweeting behavior prediction based on one-class collaborative filtering in social networks,” International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.977–980, 2016.
[10] Y. Pan, F. Cong, K. Chen, and Y. Yu, “Diffusion-aware personalized social update recommendation,” ACM Conference on Recommender Systems, pp.69–76, 2013.
[11] L.J. Hong, A.S. Doumith, and B.D. Davison, “Co-factorization machines: modeling user interests and predicting individual decisions in twitter,” the 6th ACM International Conference on Web Search and Data Mining, pp.557–566, 2013.
[12] T.R. Zaman, R. Herbrich, J.V. Gaal, and D. Stern, “Predicting information spreading in twitter,” Workshop on Computational Social Science and the Wisdom of Crowds, vol.104, no.45, pp.17599–17601, 2010.
[13] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, “Collaborative personalized tweet recommendation,” Sigir, pp.661–670, 2012.
[14] J. Zhang, J. Tang, J. Li, Y. Liu, and C. Xing, “Who influenced you? predicting retweet via social influence locality,” ACM Transactions on Knowledge Discovery from Data, vol.9, no.3, p.25, 2015.
[15] J. Bian, Y. Yang, and T.S. Chua, “Predicting trending messages and diffusion participants in microblogging network,” International
ACM SIGIR conference on Research & development in information retrieval, pp.537–546, 2014.

[16] S.A. Myers, A. Sharma, P. Gupta, and J. Lin, “Information network or social network?: the structure of the twitter follow graph,” International Conference on World Wide Web, pp.493–498, 2014.

[17] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of social networks in information diffusion,” International Conference on World Wide Web, vol.8, no.22, pp.519–528, 2012.

[18] J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li, “Social influence locality for modeling retweeting behaviors,” International Joint Conference on Artificial Intelligence, pp.2761–2767, 2013.

[19] R. Kobayashi and R. Lambiotte, “Tideh: Time-dependent hawkes process for predicting retweet dynamics,” International Conference on Web and Social Media, pp.191–200, 2016.

[20] J. Tang, S. Wu, and J. Sun, “Confluence: conformity influence in large social networks,” ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.347–355, 2013.

[21] Z. Xu and Q. Yang, “Analyzing user retweet behavior on twitter,” Ieee/acm International Conference on Advances in Social Networks Analysis and Mining, pp.46–50, 2013.

[22] H. Ma, W. Qian, F. Xia, X. He, J. Xu, and A. Zhou, “Towards modeling popularity of microblogs,” Frontiers of Computer Science, vol.7, no.2, pp.171–184, 2013.

[23] Z. Luo, M. Osborne, J. Tang, and T. Wang, “Who will retweet me?: finding retweeters in twitter,” International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.869–872, 2013.

[24] Y. Liu, S. Xu, and G. Tournass, “Detecting rumors through modeling information propagation networks in a social media environment,” International Conference on Social Computing, pp.121–130, 2015.

[25] A.H. Phan and A. Cichocki, “Tensor decompositions for feature extraction and classification of high dimensional datasets,” NOLTA, vol.1, no.1, pp.37–68, 2010.

[26] M. Wang, W. Zuo, and Y. Wang, “A multidimensional nonnegative matrix factorization model for retweet behavior prediction,” Mathematical Problems in Engineering, 2015, (2015-3-1), vol.2015, pp.1–10, 2015.

[27] B. Jiang, J. Liang, Y. Sha, and L. Wang, “Message clustering based matrix factorization model for retweeting behavior prediction,” the 24th ACM International on Conference on Information and Knowledge Management, pp.1843–1846, 2015.

[28] T.G. Kolda and J. Sun, “Scalable tensor decompositions for multi-aspect data mining,” 8th IEEE International Conference on Data Mining, pp.363–372, 2009.

[29] E.E. Papalexakis, C. Faloutsos, and N.D. Sidiropoulos, “Parcube: Sparse parallelizable tensor decompositions,” ACM Transactions on Knowledge Discovery from Data, vol.10, no.1, pp.521–536, 2012.

[30] B.Y. Zou, L.I. Cui-Ping, L.W. Tan, H. Chen, and S.Q. Wang, “Social recommendations based on user trust and tensor factorization,” Journal of Software, vol.25, no.12, pp.2852–2864, 2014.

[31] A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver, “Multiversion recommendation: n-dimensional tensor factorization for context-aware collaborative filtering,” ACM Conference on Recommender Systems, Recsys 2010, Barcelona, Spain, Sept., pp.79–86, 2010.

[32] C. Sarkar and S. Jalan, “Randomness and structure in collaboration networks: A random matrix analysis,” IEEE Trans. Comput. Social Syst., vol.3, no.3, pp.132–138, 2016.

[33] D.M. Dunlavy, T.G. Kolda, and E. Acar, “Temporal link prediction using matrix and tensor factorizations,” ACM Transactions on Knowledge Discovery from Data, vol.5, no.2, pp.1–27, 2010.

[34] G. Bergqvist and E.G. Larsson, “The higher-order singular value decomposition: Theory and an application [lecture notes],” IEEE Signal Process. Mag., vol.27, no.3, pp.151–154, 2010.

[35] M. Jenders, G. Kasneci, and F. Naumann, “Analyzing and predicting viral tweets,” International Conference on World Wide Web, pp.657–664, 2013.