Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 936397, 10 pages
http://dx.doi.org/10.1155/2015/936397

Research Article

A Multidimensional Nonnegative Matrix Factorization Model for Retweeting Behavior Prediction

Mengmeng Wang,[1,2] Wanli Zuo,[1,2] and Ying Wang[1,2,3]

[1]*College of Computer Science and Technology, Jilin University, Changchun 130012, China*
[2]*Key Laboratory of Symbolic Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun 130012, China*
[3]*College of Mathematics, Jilin University, Changchun 130012, China*

Correspondence should be addressed to Ying Wang; 726854768@qq.com

Received 22 October 2014; Accepted 6 February 2015

Academic Editor: Sergio Preidikman

Today microblogging has increasingly become a means of information diffusion via user's retweeting behavior. As a consequence, exploring on retweeting behavior is a better way to understand microblog's transmissibility in the network. Hence, targeted at online microblogging, a directed social network, along with user-based features, this paper first built content-based features, which consisted of URL, hashtag, emotion difference, and interest similarity, based on time series of text information that user posts. And then we measure relationship-based factor in social network according to frequency of interactions and network structure which blend with temporal information. Finally, we utilize nonnegative matrix factorization to predict user's retweeting behavior from user-based dimension and content-based dimension, respectively, by employing strength of social relationship to constrain objective function. The results suggest that our proposed method effectively increases retweeting behavior prediction accuracy and provides a new train of thought for retweeting behavior prediction in dynamic social networks.

1. Introduction

With the development of Internet, microblogging has become a novel social media [1, 2]. As a platform, which shares, disseminates, and accesses information based on relationships between users, microblogging is originality, timeliness, grassroots, randomness, debris, and so forth. It is not only a tool for communication and self-expression, but also a means of information releasing and public relations marketing for governments, enterprises, and organizations. The emergence of microblogging greatly speeded up transmission of information in the network. For an instance, in order to share information with friends, a user can quickly copy information which he/she is interested in to his/her own microblogging space through retweeting function which provides a convenient way to air user's opinion, as well as a way for users to communicate with each other. Thanks to its great flexibility, this way of information transmission is favored by communicators; meanwhile, it brings a viral spread of a microblog through retweeting behavior between different users for its being nonmandatory, targeted and personalized.

Moreover, social shared content (such as retweeting a microblog) is not random but depends on the transmissibility of its own. As a consequence, exploring on microblogging retweeting behavior can make us better understand diffusion of information in the network, as well as help identify information credibility [3], reorder user's tweets [4], and identify interesting tweet [5]. And it can also be used to recommend microblogs to users according to their interested topics which may be reflected in their retweeting microblogs. Furthermore, research has shown that users were more inclined to share contents that can stimulate their emotions [6]. Hence, retweeting behavior prediction is of great significance for emotion analysis and public opinion monitoring. Our work on predicting user's retweeting behavior is motivated by its broad application prospect.

At present, users' propensities on retweeting remain unclear. Retweeting behavior prediction is in the stage of development; consequently, there are still some unsolved

problems in this field. To this end, we proposed a multidimensional nonnegative matrix factorization model for retweeting behavior prediction in social networks (denoted as MNMFRP), and our main contributions are summarized next.

(1) Different from previous methods which predicted retweeting behavior without taking user's emotions into consideration, we put forward a new concept, emotion difference, which represented difference between the emotion reflected in user's recent contents and a certain microblog's sentiment. And along with URL, hashtag, and interest similarity, emotion difference was regarded as a content-based factor in the problem of retweeting behavior prediction so as to gain performance.

(2) In order to be applied to dynamic networks better, on the basis of time series of user's contents and user's network topological information, we considered network as a dynamic flow of time slices and distributed different weights to different time slices to blend temporal information in retweeting behavior prediction algorithm, so as to capture dynamic evolution process of information and network structure.

(3) Traditional Salton metrics which calculated adjacent degree between nodes based on network topological information was appropriate for undirected networks; according to directivity of link, we improved traditional Salton metrics and combined it with frequency of interactions to measure relationship-based factor in a social network.

(4) Employing strength of social relationship to constrain objective function, we cast the predicting problem into solutions of nonnegative matrix factorization from user-based dimension and nonnegative matrix factorization from content-based dimension, respectively, so as to reduce complexity effectively.

The rest of paper is organized as follows: Section 2 describes the related work; Section 3 defines the method we propose; details of the experimental results and dataset which is used in this study are given in Section 4. Finally conclusion appears in Section 5.

2. Related Work

Since Twitter appeared in 2005, the influence of microblogging in society has increased. And making full use of user's behavior for precision marketing is the target of each enterprise [7]. Since the problem of retweeting behavior prediction which has become a hotspot in recent years is the key to understand how information transmits in the network, hence, more and more scholars shifted emphasis towards how to mine characteristics of user's retweeting behavior and carried on thorough researches on it which achieved some progress.

In terms of retweeting function of Twitter, Boyd et al. [8] made a detailed analysis to explore how people retweet, why people retweet, and what people retweet. Yang et al. [9] analyzed influence that user, information, and time had on user's retweeting behavior in Twitter. Based on their observation, they put forward a factor graph model to predict user's retweeting behavior and the spread of a new microblog. Zaman et al. [10] divided user's characteristics in Twitter into three categories: characteristics of publisher (nickname, the count of followers, and the count of followees), characteristics of retweeter (nickname, the count of followers, and the count of followees) and characteristics of content (length of content), and predicted retweeting probability of a single user via Matchbox [11]; nevertheless, they did not take the correlation between user's interests and content of microblog into account; instead, they simply analyzed inherent characteristics of microblog and user individually. Suh et al. [12] first collected a large amount of Twitter data and extracted a set of semantic irrelevant attributes (the number of links, tags, @usernames, followers, followees, historical contents and created time of account, and so on). And then they established retweeting behavior prediction model based on Generalized Linear Model (denoted as GLM). They found that the number of links, tags, @usernames, followers, followees, and created time of account had impacts on retweeting rate of a microblog, while the number of historical contents had nothing to do with retweeting rate of a microblog. However, they only considered characteristics of single node and text, without considering the effect that relationship between nodes had on retweeting rate of a microblog. Through simulating a retweeting behavior curve which represented the number of retweets in a successive time unit, Zhang et al. [13] predicted the number of retweets on the basis of events. Petrovic et al. [14] leveraged an improved passive-aggressive algorithm to predict retweeting behavior which can merely predict 46.6% of microblogs accurately.

Furthermore, a great many of studies have shown that contents [15–17] and emotions [18, 19] of microblogs played very important roles in information transmission. Tan et al. [20] defined three factors: behavior preference factor, friends influence factor, and self-related factor, and given to their impacts on user's retweeting behavior, they put forward a noise tolerant time-varying factor graph model (denoted as NTT-FGM) to simulate and predict user's retweeting behavior in social network which converted predicting problem into a conditional probability problem to solve. By utilizing user's social network structure, linguistic features derived from user's historical tweets, Artzi et al. [21] predicted retweeting probability with Multiple Additive Regression Trees (MART) and a maximum entropy classifier which not only provided a fast classification, but also met a natural requirement for large scale real-time tasks. The results revealed that the proposed methods were capable of generating accurate prediction for a large number of tweets. Bandari et al. [22] put forward an algorithm to predict whether a piece of news could be popular on Twitter or trigger a heated discussion on social network sites. They explored the relations between the amount of retweeting and their proposed features which included category of content, objective degree, person, and place which were mentioned and source of content via regression algorithm. In order to predict retweeting behavior, they divided popularity of microblog into three levels according to retweeting quantity of a microblog. Although prediction accuracy of each level reached 84%, the proposed algorithm could merely infer

the amount of retweeting after being shared on the basis of microblogs' contents. Hong et al. [23] explored factors that influenced information propagation in Twitter based on content of message, temporal information, metadata of message, and user, as well as structural properties of user's social graph on a large scale dataset. Through predicting whether or not a message will be retweeted and the volume of retweets a particular message will receive in the near future, they successfully predicted popularity of messages with good performance. Bae et al. [24] extracted tweets with similar topics, retweeting patterns, and properties based on social, local, and content features to make a prediction which achieved significant results. Tsur and Rappoport [25] combined content features with temporal and topological features to predict the spread of an idea in a given time frame via a hybrid approach based on a linear regression. Xu and Yang [26] performed a general analysis of retweeting behavior from the perspective of individual users via J48, SVM, and logistic regression and made "leave-one-feature-out" comparisons of four different types of features: social-based, content-based, tweet-based, and author-based features which demonstrated social-based, content-based features were strongly associated with user's retweeting behavior. Zhang et al. [27] demonstrated the existence of influence locality in microblogging network and predicted user's retweeting behavior based on social influence locality via a logistic regression classifier which can obtain a $F1$-measure of 71.65% without any additional features and a $F1$-measure of 73.3% with personal attributes, instantaneity, topic propensity, and social influence locality.

Nevertheless, research on retweeting behavior prediction in dynamic social networks is still in its infancy; many issues and research methods are derived from static networks and undirected networks, consequently, there are still some unsolved problems in this field. In this paper, through carrying on detailed analysis of directivity of link, impacts that attributes of users, contents, and relationships had on retweeting behavior, dynamic evolution process of network, and fusion of multidimensional factors, we presented a multidimensional nonnegative matrix factorization model for retweeting behavior prediction in order to improve the accuracy of retweeting behavior prediction in dynamic and directed social networks.

3. Multidimensional Nonnegative Matrix Factorization Model for Retweeting Behavior Prediction

Since factors which can influence a user to retweet blogs of other users are very few, furthermore, retweeting matrix is very sparse and low-rank; consequently, users and contents can have a more compact but accurate representation in a low-rank space. Therefore, in this section, we first defined factors that had impact on retweeting behavior, and then models on user-based features and content-based features were constructed via nonnegative matrix factorization, followed by fusing models according to their error rates.

3.1. Features of Retweeting Behavior Prediction. In social networks, there are two kinds of relationships between users, namely, following and being followed. Microblogs released by followers spread in the network mainly through retweeting behavior of followees; thus, retweeting behavior is the main channel of information dissemination; it also reflects users' individual difference. Therefore, it is important to analyze factors that affect user's retweeting behavior which is the basis of establishing prediction model. Whether a user will retweet a microblog or not is mainly decided by three factors: characteristics of microblog [15–17], characteristics of user, and characteristics of relationship. As a consequence, we fully took advantage of information in user's profile and contents that user released to extract useful characteristics for retweeting behavior prediction.

3.1.1. User-Based Features. Users who retweet different categories of microblogs may have different genders. For instance, users who retweet microblogs about beauty and gossip are mostly females, while a large proportion of users who retweet microblogs about sports and digital technology are males. So user's gender has significant impact on whether a user will retweet a certain type of microblog. Additionally, the number of user's followers, followees, and contents that user post may reflect user's retweeting possibility as well. And in this paper, we constructed user-based features which consisted of the number of bifollowers, followers and followees, contents that user posted, user's province and city, user's gender, created time, and verified type of user's account.

3.1.2. Content-Based Features

URL and Hashtag. Researchers have shown that 30% microblogs which were retweeted a lot of time contained URLs [12]; URL was an important factor in retweeting behavior prediction [28, 29]. Hence, we took URL as a feature in our framework; if a microblog contains URLs, then its URL value is 1, and 0 otherwise. Besides, hashtag was also taken into consideration in this work; if a microblog contains hashtags, then its hashtag value is 1, and 0 otherwise.

Emotion Difference. Moreover, studies have shown that user's emotion played a very important role in information transmission [18, 19], emotional resonance between users and microblogs made users easier to retweet microblogs. Thus, to enhance the performance of retweeting behavior prediction, we conducted sentiment analysis on user's contents with the corpus of HowNet Knowledge (http://www.keenage.com/download/sentiment.rar). HowNet Knowledge, which includes 8945 words and phrases, consists of six files: positive emotional words list file, negative emotional words list file, positive review words list file, negative review words list file, degree words list file, and propositional words list file. According to the results of sentiment analysis, we obtained a user's and a microblog's emotion, and then given dynamic nature of social network, we considered network as a dynamic flow of time slices and summed up user's emotion

of different time slices with different weights to integrate user's emotion with temporal information.

First, we calculated user's emotion expressed in his/her contents on the ith time slice t_i as follows:

$$Ue(u,t_i) = \frac{Upn(u,t_i)}{Unn(u,t_i)}, \tag{1}$$

where $Ue(u,t_i)$ represents the emotion of user u's contents on the ith time slice t_i; $Upn(u,t_i)$ and $Unn(u,t_i)$ represent the number of positive emotional words and the number of negative emotional words user u uses on the ith time slice t_i which are included in HowNet Knowledge. Thus, user's recent emotion on the flow of time slices $[0,t_n]$ is calculated with

$$Ue^{[0,t_n]}(u) = \sum_{i=0}^{n} \alpha^{n-i} \times Ue(u,t_i), \tag{2}$$

where α^{n-i} represents the weight of the ith time slice t_i. And then we calculated the emotion of a microblog as

$$Ie(b) = \frac{Ipn(b)}{Inn(b)}, \tag{3}$$

where $Ie(b)$ represents the emotion of microblog b; $Ipn(b)$ and $Inn(b)$ represent the number of positive emotional words and the number of negative emotional words used in microblog b which are included in HowNet Knowledge. So emotion difference between user u and microblog b is calculated as

$$Em(u,b,t_n) = \left| Ue^{[0,t_n]}(u) - Ie(b) \right|. \tag{4}$$

Interest Similarity. Webberley et al. [30] pointed out that users tended to post or retweet a microblog which he was interested in. Furthermore, user's interests can be reflected in the content of user's microblog, and users who retweet similar kinds of microblogs may have similar interests. Therefore, we first extracted user's interests from microblogs that user released and extracted topics from a certain microblog simultaneously. Then we calculated the similarity between user's interest set and a certain microblog's topic set which is shown in

$$Is(u,b) = \frac{|it(u) \cap t(b)|}{|it(u) \cup t(b)|}, \tag{5}$$

where u denotes a user, b denotes a certain microblog, $it(u)$ denotes interest set of u, and $t(b)$ denotes topic set of b.

3.1.3. Relationship-Based Feature

Strength of Social Relationship. Similar to human society, a user who has intimate relations with others will be more likely to get approval from others, so that his/her posts are easier to be retweeted. Besides, Yang and Rim [5] pointed out that propagation of microblogs might be limited to user's social network. As network topology can reflect user's status in social network implicitly, some researchers held the view that the more similar users' status was, the more similar their network topological structures were, and the stronger social relationship was between them. Additionally, the more frequently users interact with each other, the stronger social relationship they may have with each other. Hence, in this paper, we inferred strength of social relationship from users' network topological structures and frequency of interactions. Since relationships would decay with time [31], we combined related factors with temporal information and put forward dynamic factors to measure strength of social relationship in social network effectively.

Common neighbors metrics assumed that similarity between users was proportional to the number of their common neighbors. In this paper, we adopted Salton metrics which introduced users' degree compared to common neighbors metrics to measure user's network topological structure. In the same way with emotion difference, first of all, based on directivity of link, we made an improvement on traditional Salton metrics and Salton metrics of user u and user v on the ith time slice t_i is defined as follows:

$$Sa(u,v,t_i) = \frac{\left|\Gamma^{\text{in}}(u,t_i) \cap \Gamma^{\text{in}}(v,t_i)\right| / \sqrt{d^{\text{in}}(u,t_i) \times d^{\text{in}}(v,t_i)}}{\left|\Gamma^{\text{out}}(u,t_i) \cap \Gamma^{\text{out}}(v,t_i)\right| / \sqrt{d^{\text{out}}(u,t_i) \times d^{\text{out}}(v,t_i)}}, \tag{6}$$

where $\Gamma^{\text{in}}(u,t_i)$ and $\Gamma^{\text{in}}(v,t_i)$ stand for in-link users set of user u and user v on the ith time slice t_i, respectively, $\Gamma^{\text{out}}(u,t_i)$ and $\Gamma^{\text{out}}(v,t_i)$ stand for out-link users set of user u and user v on t_i, respectively, in-link and out-link are defined by follower relationship, $|\Gamma(x)|$ stands for the number of elements in set $\Gamma(x)$, $d^{\text{in}}(u,t_i)$ and $d^{\text{in}}(v,t_i)$ represent in-degree of user u and user v on t_i, respectively, and $d^{\text{out}}(u,t_i)$ and $d^{\text{out}}(v,t_i)$ represent out-degree of user u and user v on t_i, respectively. Thus, Salton metrics of user u and user v on the flow of time slices $[0,t_n]$ is calculated as

$$Sa^{[0,t_n]}(u,v) = \sum_{i=0}^{n} \alpha^{n-i} \times Sa(u,v,t_i). \tag{7}$$

Similarly, frequency of interaction between user u and user v on the ith time slice t_i is defined as follows:

$$\text{Fre}(u,v,t_i) = \frac{c(u,v,t_i)}{p(u,t_i) + p(v,t_i)}, \tag{8}$$

where $p(u,t_i)$ and $p(v,t_i)$ stand for the number of posts from user u and user v on t_i, respectively. $c(u,v,t_i)$ stands for the number of interactions between user u and user v on t_i where an interaction is defined as a user retweeting a microblog of another user. Thus, frequency of interactions between user u and user v on the flow of time slices $[0,t_n]$ is calculated as

$$\text{Fre}^{[0,t_n]}(u,v) = \sum_{i=0}^{n} \alpha^{n-i} \times \text{Fre}(u,v,t_i). \tag{9}$$

Finally, strength of social relationship between user u and user v is calculated as follows:

$$S(u,v) = \beta \times Sa^{[0,t_n]}(u,v) + (1-\beta) \times \mathrm{Fre}^{[0,t_n]}(u,v), \quad (10)$$

where β is employed to control the contribution of each factor.

3.2. Algorithm of Multidimensional Nonnegative Matrix Factorization Model for Retweeting Behavior Prediction. Let $u = \{u_1, u_2, \ldots, u_n\}$ be users set where n denotes the number of users and let $b = \{b_1, b_2, \ldots, b_m\}$ be blogs set where m denotes the number of microblogs. $R \in \mathbb{R}^{n \times m}$ is user-blog retweeting matrix where $R_{ij} = 1$ if the ith user u_i retweets the jth blog b_j and $R_{ij} = 0$ otherwise.

We first factorized R into matrix $U_1 \in \mathbb{R}^{n \times d_1}$ and matrix $V_1 \in \mathbb{R}^{m \times d_1}$ where U_1 denotes user-based features matrix, $d_1 \ll n, m$ denotes the number of user-based features, and V_1 captures the correlations between R's and U_1's low-rank representations. Initially, on the basis of user-based features, we minimized the squared error between predicted retweeting behavior and observed retweeting behavior as follows:

$$\min_{U,V} \left\| R - U_1 V_1 \right\|_F^2, \quad (11)$$

where $\| \cdot \|_F$ is the Frobenius norm fitting constraint. Moreover, to avoid overfitting, on the basis of (11), we added Frobenius regularization norm on U_1 and V_1, respectively:

$$\min_{U,V} \left\| R - U_1 V_1 \right\|_F^2 + \lambda_1 \left\| U_1 \right\|_F^2 + \lambda_2 \left\| V_1 \right\|_F^2, \quad (12)$$

where λ_1 and λ_2 are regularization parameters. Furthermore, we added social relationship regularization terms into (12) to constrain personality difference between one user and another on account of the assumption that the stronger the social relationship two users have, the smaller the difference between them in both their preferences and personalities. As long as we constrain personality difference between users to be small, we may arrive at the following social relationship Laplacian regularizer:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} S(i,j) \left\| (U_1)_{i*} - (U_1)_{j*} \right\|_F^2 = \mathrm{Tr}\left(U_1^{\mathsf{T}} \mathscr{L} U_1\right), \quad (13)$$

where $S(i,j) \in [0,1]$ stands for the strength of social relationship between u_i and u_j, and we assumed that the larger $S(i,j)$ is, the more likely u_i may retweet microblogs of u_j, while a smaller value of $S(i,j)$ tells that the distance of their latent representations should be larger. $(U_1)_{i*}$ and $(U_1)_{j*}$ represent the ith and the jth sample of U_1, respectively. $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix and $\mathscr{L}$ denotes the Laplacian matrix which is calculated as $\mathscr{L} = D - \mathscr{Z}$, where D represents a diagonal matrix in which the ith element $D(i,i)$ equals the sum of the ith row vector of S. Hence, we obtained the following objective function through imposing the above social relationship Laplacian regularizer onto (12):

$$\min_{U,V} F_1 = \left\| R - U_1 V_1 \right\|_F^2 + \lambda_1 \left\| U_1 \right\|_F^2 + \lambda_2 \left\| V_1 \right\|_F^2 + \lambda_3 \mathrm{Tr}\left(U_1^{\mathsf{T}} \mathscr{L} U_1\right), \quad (14)$$

where if V_1 is fixed, F_1 is a convex optimization problem with respect to U_1 and if U_1 is fixed, F_1 is a convex optimization problem with respect to V_1. However, when both U_1 and V_1 are not fixed, F_1 is not convex. Therefore, global optimal solutions of F_1 is difficult to formalize. Nevertheless, local optimal solution of F_1 can be obtained by multiplicative update method [32].

In order to obtain the updating rules of U_1 and V_1, after removing constants in the objective function, we defined the Lagrangian function of (14) which is shown as follows:

$$\mathscr{L}_{F_1} = \mathrm{Tr}\left((R - U_1 V_1)^2\right) + \lambda_1 \mathrm{Tr}\left(U_1 U_1^{\mathsf{T}}\right) + \lambda_2 \mathrm{Tr}\left(V_1 V_1^{\mathsf{T}}\right) + \lambda_3 \mathrm{Tr}\left(U_1^{\mathsf{T}} \mathscr{L} U_1\right) - \mathrm{Tr}\left(\psi U_1\right) - \mathrm{Tr}\left(\varphi V_1\right), \quad (15)$$

where ψ and φ are Lagrangian multipliers for nonnegativity of U_1 and V_1, respectively. And then we took the gradient of (15) with respect to U_1 and V_1 and setting them to zero, respectively:

$$\frac{\partial \mathscr{L}_{F_1}}{\partial U_1} = -2\left(RV_1\right)^{\mathsf{T}} + 2U_1 V_1 V_1^{\mathsf{T}} + 2\lambda_1 U_1 + 2\lambda_3 (D - \mathscr{Z}) U_1 - \psi = 0$$
$$\frac{\partial \mathscr{L}_{F_1}}{\partial V_1} = -2U_1^{\mathsf{T}} R + 2U_1^{\mathsf{T}} U_1 V_1 + 2\lambda_2 V_1 - \varphi = 0. \quad (16)$$

By multiplying (16) by U_1 and V_1, (16) can be written as

$$-2\left(RV_1\right)^{\mathsf{T}} U_1 + 2U_1 V_1 V_1^{\mathsf{T}} U_1 + 2\lambda_1 U_1 U_1 + 2\lambda_3 (D - \mathscr{Z}) U_1 U_1 - \psi U_1 = 0$$
$$-2U_1^{\mathsf{T}} R V_1 + 2U_1^{\mathsf{T}} U_1 V_1 V_1 + 2\lambda_2 V_1 V_1 - \varphi V_1 = 0. \quad (17)$$

Given KKT (Karush-Kuhn-Tucker) optimal conditions, $\psi U_1 = 0$ and $\varphi V_1 = 0$, we can get the following equation:

$$-2\left(RV_1\right)^{\mathsf{T}} U_1 + 2U_1 V_1 V_1^{\mathsf{T}} U_1 + 2\lambda_1 U_1 U_1 + 2\lambda_3 (D - \mathscr{Z}) U_1 U_1 = 0$$
$$-2U_1^{\mathsf{T}} R V_1 + 2U_1^{\mathsf{T}} U_1 V_1 V_1 + 2\lambda_2 V_1 V_1 = 0, \quad (18)$$

where each element in R, D, and $\mathscr{Z}$ is nonnegative and λ_1, λ_2, and λ_3 are nonnegative as well, moreover, the initial values of U_1 and V_1 are both nonnegative; consequently, $(RV_1)^{\mathsf{T}}$, $\lambda_3 \mathscr{Z} U_1$, $U_1 V_1 V_1^{\mathsf{T}}$, $\lambda_1 U_1$, $\lambda_3 D U_1$, $U_1^{\mathsf{T}} R$, $U_1^{\mathsf{T}} U_1 V_1$, and $\lambda_2 V_1$ are

Table 1: Descriptions of files in SinaWeibo dataset.

File name	Description
weibo_network.txt	This file describes the SinaWeibo subnetwork at the very first timestamp, namely, static following network
graph_170w_1month.txt	This file describes the SinaWeibo subnetwork at every timestamp (one timestamp is one day) in one month from 2012.9.28 to 2012.10.29, namely, dynamic following network
user_profile1.txt and user_profile2.txt	These files describe users' profile
root_content.txt	This file describes original tweet content
uidlist.txt	This file maps the original user ID to the new ID

nonnegative. Thus (18) can be written as follows in which both ends are nonnegative:

$$\begin{aligned}\left(U_1V_1V_1^{\intercal} + \lambda_1 U_1 + \lambda_3 DU_1\right)U_1 &= \left(\left(RV_1\right)^{\intercal} + \lambda_3 \mathcal{Z}U_1\right)U_1\\ \left(U_1^{\intercal}U_1V_1 + \lambda_2 V_1\right)V_1 &= \left(U_1^{\intercal}R\right)V_1.\end{aligned} \tag{19}$$

So the updating rules of U_1 and V_1 are defined as follows:

$$\begin{aligned}U_1 &\longleftarrow U_1 \frac{\left(RV_1\right)^{\intercal} + \lambda_3 \mathcal{Z}U_1}{U_1V_1V_1^{\intercal} + \lambda_1 U_1 + \lambda_3 DU_1}\\ V_1 &\longleftarrow V_1 \frac{U_1^{\intercal}R}{U_1^{\intercal}U_1V_1 + \lambda_2 V_1}.\end{aligned} \tag{20}$$

Similarly, the objective function based on content-based features is shown as follows:

$$\min_{U,V} F_2 = \left\|R - U_2V_2\right\|_F^2 + \lambda_4 \left\|U_2\right\|_F^2 + \lambda_5 \left\|V_2\right\|_F^2, \tag{21}$$

where R is factorized into matrix $U_2 \in \mathbb{R}^{n\times d_2}$ and $V_2 \in \mathbb{R}^{m\times d_2}$, where U_2 denotes content-based features matrix, $d_2 \ll n, m$ denotes the number of content-based features, and V_2 captures the correlations between R's and U_2's low-rank representations. λ_4 and λ_5 are regularization parameters. And the updating rules of U_2 and V_2 are defined as follows:

$$\begin{aligned}U_2 &\longleftarrow U_2 \frac{\left(RV_2\right)^{\intercal}}{U_2V_2V_2^{\intercal} + \lambda_4 U_2}\\ V_2 &\longleftarrow V_2 \frac{U_2^{\intercal}R}{U_2^{\intercal}U_2V_2 + \lambda_5 V_2}.\end{aligned} \tag{22}$$

Based on the above analysis, we summarize the detailed algorithm in Algorithm 1.

3.3. Time Complexity. Assume that the number of iterations is T. The complexity of multidimensional nonnegative matrix factorization model for retweeting behavior prediction is analyzed as follows. The updating rules of U_1, V_1, U_2, and V_2, from step (10) to step (15), contribute most to the time complexity of the proposed algorithm. Since both R and S are very sparse, $(RV_1)^{\intercal}$ and $U_1V_1V_1^{\intercal}$ take $O(nmd_1)$ and $O(\max(n,m){d_1}^2)$ time, respectively. In addition, D is a diagonal matrix; the time complexity of DU_1 is $O(nd_1)$. As $d_1 \ll \min(n,m)$, thus the updating rule for U_1 is taking $O(nmd_1)$ time. Similarly, the updating rule of U_2 can be computed in $O(nmd_2)$ time. Hence, the overall complexity for our proposed method is $T \times O(nm(d_1 + d_2))$.

4. Experimental Evaluation

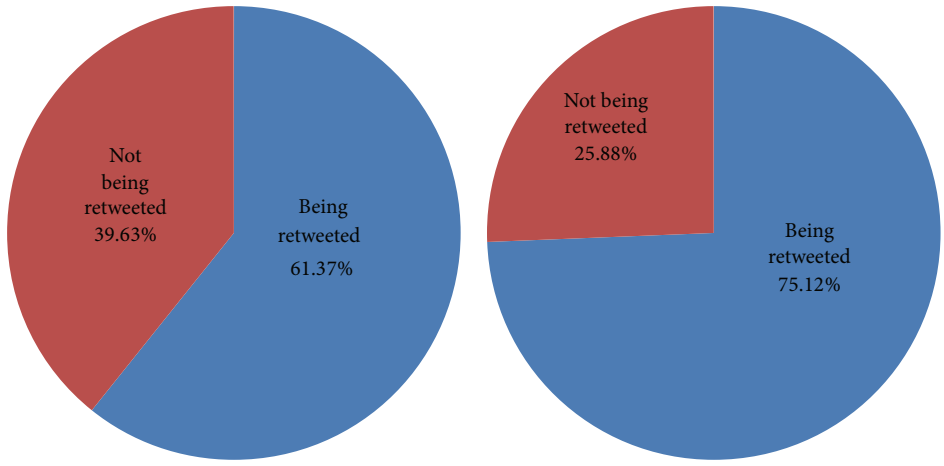
4.1. Dataset. So far, microblogging has become one of important sources of information, as well as a main channel of information dissemination. Governments, enterprises, and public figures use microblog for marketing or guiding public opinion. "2013 research report of user behavior on China social application," which was released by China Internet Network Information Center in November 2013, showed the coverage of microblogging users has reached 55.4% in 2013 after its explosive growth in three years. Therefore, we leveraged Sina microblog dataset [27] to evaluate validity of the method we proposed; documents of Sina microblog dataset are shown in Table 1.

4.2. Analysis of Retweeting Behavior Prediction Features. First of all, we carried on an analysis of URL and hashtag in microblogs. There are 61819 microblogs that contain URLs and 36545 microblogs that contain hashtags. The retweeting proportion of microblogs that contains these characteristics is shown in Figure 1. As it can be seen from Figure 1, quite amount of microblogs that contain URLs and hashtags are being retweeted which demonstrated that URLs and hashtags played important roles in retweeting behavior prediction [12, 28, 29]. Meanwhile, we explored the effects that proposed emotion difference, interest similarity, and strength of social relationship had on retweeting probability which are shown in Figure 2.

Figure 2(a) shows strong evidence for the impact that emotion discrepancies has on user's retweeting behavior. The ratio of retweeted microblogs is higher if there is small difference between microblog's emotion and emotion expressed in user's recent states; this phenomenon might explain that the smaller the emotion discrepancies is, the more it is easy to resonate in emotion, so does the possibility of retweeting increase. And we also find that users will have higher probability to retweet if there is significant large difference between microblog's emotion and emotion expressed in user's recent states. This is due to the reason that if a user suffers from poor emotional state lately, when he/she encounters with a microblog whose emotion is far

Input: retweeting matrices R; user-based features matrix U_1; content-based features matrix U_2; parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$
Output: predicted retweeting matrix R'
The description of the algorithm:
(1) for each $i \in \{1, 2, \ldots, n\}$ do
(2) for each $j \in \{1, 2, \ldots, n\}$ do
(3) if $R_{ij} = 1$ or $R_{ji} = 1$ then
(4) calculate $S(i, j)$ with (10)
(5) end if
(6) end for
(7) end for
(8) Initialize V_1 with $(U_1^T U_1)^{-1} R$ and set elements which are negative in V_1 to 0
(9) Initialize V_2 with $(U_2^T U_2)^{-1} R$ and set elements which are negative in V_2 to 0
(10) Repeat
(11) update U_1 and V_1 with (20)
(12) Until F_1 in (14) reaches to convergence
(13) Repeat
(14) update U_2 and V_2 with (22)
(15) Until F_2 in (21) reaches to convergence
(16) Calculate error rate of F_1: $e_1 = |\text{err}_1|/m$
% $|\text{err}_1|$ denotes the number of instances which are wrongly classified with F_1
(17) Calculate error rate of F_2: $e_2 = |\text{err}_2|/m$
% $|\text{err}_2|$ denotes the number of instances which are wrongly classified with F_2
(18) Return $R' = (e_1 e_2 / e_1(e_1 + e_2)) \times U_1 V_1 + (e_1 e_2 / e_2(e_1 + e_2)) \times U_2 V_2$

Algorithm 1: Multidimensional nonnegative matrix factorization model for retweeting behavior prediction.



(a) URL (b) Hashtag

Figure 1: Retweeting proportion of microblogs that contains URL and hashtag.

more different from himself/herself, he/she may tend to draw positive energy from this microblog to alleviate his/her own negative emotion; on the contrary, if a user with positive emotional state comes across a microblog that expresses negative emotion, he/she may tend to share in their suffering to appease mood of others. As in [6], the results suggested that users were more inclined to share contents that can stimulate their emotions.

And in Figure 2(b), users will have higher probability to retweet if there is larger similarity between user's interest and topics in microblog since users tend to post or retweet a microblog which they are interested in [30]. Besides, in terms of strength of social relationship, similar to human society, relation distance between users in microblogging network is also existent. As described in Figure 2(c), the stronger the social relationship is, the higher the retweeting probability is which is in keeping with Yavas and Yücel's work [33] in which they indicated that homophily had a positive effect on the extent of diffusion.

To sum up, it can be found that the proposed factors can be used as a good indicator of retweeting behavior prediction.

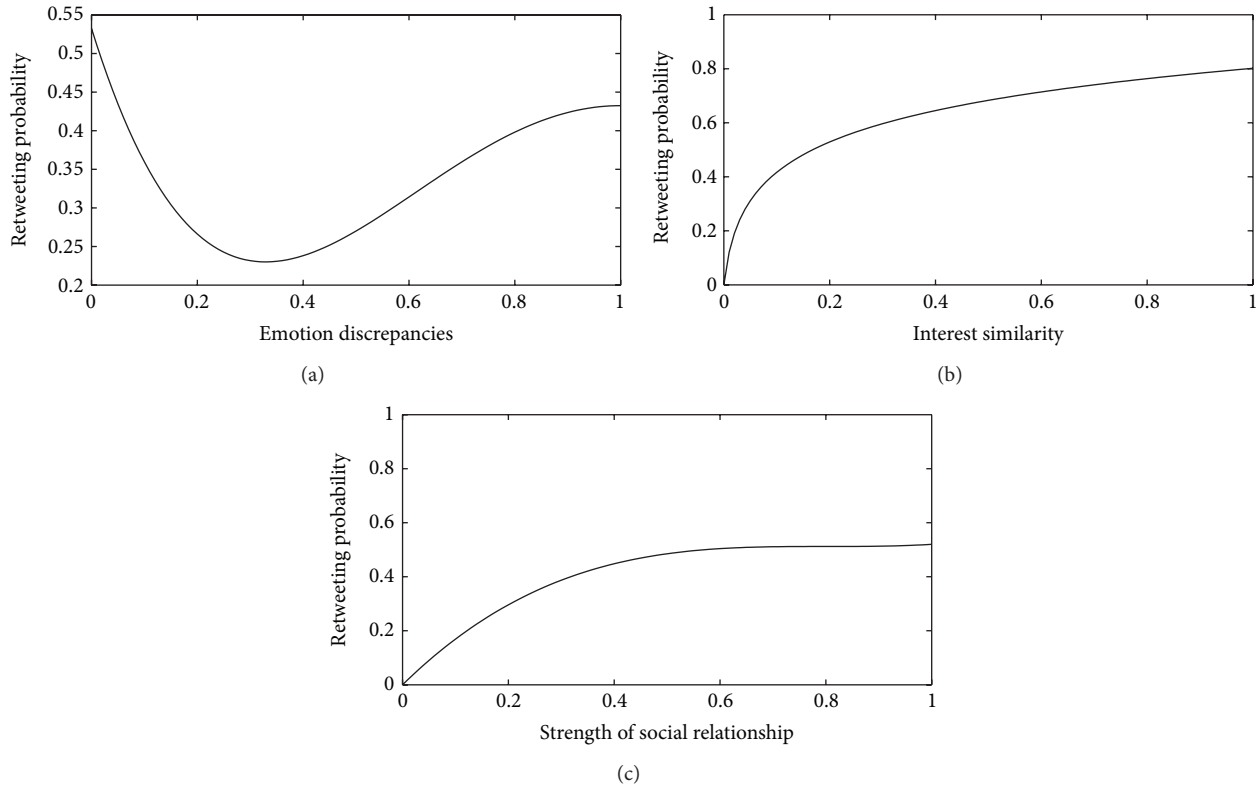


Figure 2: The relationship between retweeting probability and proposed factors.

Table 2: Results of MNMFRP on user-based features.

	Precision	Recall	$F1$-measure
Yes	0.687	0.728	0.707
No	0.484	0.434	0.457
Average	0.612	0.62	0.615

Table 3: Results of MNMFRP on content-based features.

	Precision	Recall	$F1$-measure
Yes	0.819	0.759	0.787
No	0.634	0.713	0.671
Average	0.75	0.742	0.745

4.3. Comparison of Experiments. In this section, we carried on "leave-one-feature-out" experiments on different features set, as well as experiments in different baseline methods which were evaluated via precision, recall, and $F1$-measure.

4.3.1. Experiments on Different Features Set. Traditional retweeting behavior prediction methods merely took advantage of user's static characteristics; unlike previous methods, we introduced dynamic features to enhance the performance of our method. In order to better understand the contribution of each features set, we conducted 10-fold cross validation on user-based features, content-based features, and all features with MNMFRP separately; the precision, recall and $F1$-measure of different feature sets are shown in Tables 2, 3, and 4.

By observing the results of Tables 2 and 3, removing user-based features lowers the model's prediction abilities, although prediction quality remains relatively high. Removing content-based features may create a bigger drop in performance. Hence, we find that it is difficult to predict user's retweeting probability only with limited number of user-based features. Besides, it confirms the idea in [5, 26] that it is important to consider content-based features rather than user-based features. In conclusion, removing either feature set may degrade prediction performance. As it is shown in Table 4, proposed framework that fuses multidimensional features to predict retweeting behaviors makes up for the deficiencies of traditional retweeting behavior prediction algorithms and improves accuracy of retweeting behavior prediction algorithm effectively. Additionally, by shifting the emphasis towards content-based features and strength of social relationship, we can obtain a significant improvement on performance (+9.71% in terms of precision, 2.04% in terms of recall and +6% in terms of $F1$-measure) compared with [27] which leveraged a logistic regression classifier to predict user's retweeting behavior based on the same original dataset as us.

4.3.2. Experiments in Other Baseline Methods. Finally, we leveraged other baseline methods, such as SVM, Native Bayes, BP neural network, Decision Tree, and Random Forest, to complete the prediction task on our proposed

Table 4: Results of MNMFRP based on all features.

	Precision	Recall	$F1$-measure
Yes	0.854	0.806	0.829
No	0.698	0.765	0.73
Average	0.796	0.791	0.793

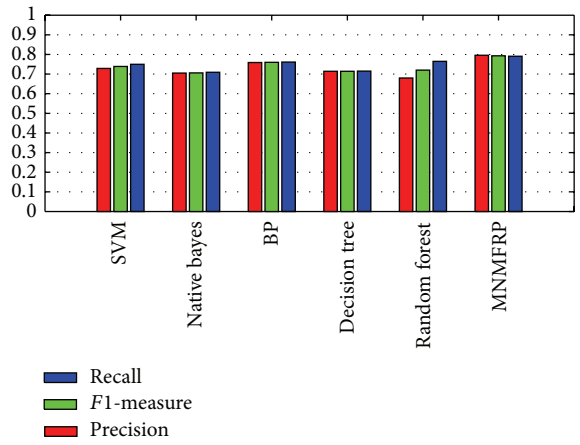


Figure 3: Average precision, recall, and $F1$-measure of different methods.

features. Due to space restrictions, Figure 3 merely shows average precision, recall, and $F1$-measure for each baseline method along with proposed method in this paper.

Besides, in order to figure out which method is more suitable for retweeting behavior prediction problem, we further evaluated execution time of different methods along with our method. Figure 4 shows the execution time, in seconds, used by each method for retweeting behavior prediction. As it can be observed from Figures 3 and 4, all measures need more time since processing of dynamic evolution in the network. The fastest measure is Native Bayes since it completes prediction task without training model; however, it performs poorly on our dataset. BP neural network achieves the better performance but is the most time-consuming measure. The next fastest measure is the method we proposed. From the above, compared to other baseline methods, our method can achieve the best performance within a shorter execution time by casting the predicting problem into solutions of nonnegative matrix factorization from user-based dimension and nonnegative matrix factorization from content-based dimension, respectively.

5. Conclusion

Aiming at the deficiencies of traditional retweeting behavior prediction algorithms, we put forward a multidimensional nonnegative matrix factorization model for retweeting behavior prediction. Firstly, on the basis of time series of users' status, we built content-based features. Secondly, relationship-based features were inferred from users' network topological structures and frequency of interactions. Finally, a retweeting behavior prediction model was proposed

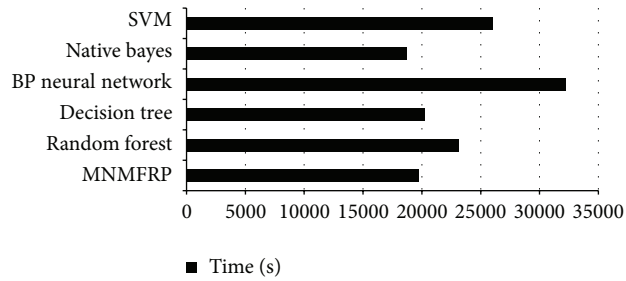


Figure 4: Execution time, in seconds, of different methods.

based on multidimensional nonnegative matrix factorization which leveraged strength of social relationship to constrain objective function. The experimental results revealed that the proposed method can effectively improve performance of retweeting behavior prediction model with collaborative features.

In future work, we will explore combining time series analysis with retweeting behavior predicting algorithm to further understand the effects that temporal information has on the networks. Additionally, we will introduce crowd sourcing technology into our method to gain performance. Furthermore, we will speculate on what directions can be undertaken to ameliorate its performance with respect to time complexity.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant no. 61300148; the Scientific and Technological Break-Through Program of Jilin Province under Grant no. 20130206051GX; the Science and Technology Development Program of Jilin Province under Grant no. 20130522112JH; the Science Foundation for China Postdoctor under Grant no. 2012M510879; the Basic Scientific Research Foundation for the Interdisciplinary Research and Innovation Project of Jilin University under Grant no. 201103129.

References

[1] H. Kwak, C. Lee, H. Park, and S. B. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 591–600, April 2010.

[2] D. Y. Zhang and G. Guo, "A comparison of online social networks and real-life social networks: a study of Sina Microblogging," *Mathematical Problems in Engineering*, vol. 2014, Article ID 578713, 6 pages, 2014.

[3] S. Sujoy, K. Byungkyu, O. John, and T. Höllerer, "Understanding information credibility on twitter," in *Proceedings of the ASE/IEEE International Conference on Social Computing*, pp. 19–24, 2013.

[4] K. Shen, J. Wu, Y. Zhang et al., "Reorder user's tweets," *ACM Transactions on Intelligent Systems and Technology*, vol. 4, no. 1, pp. 1–17, 2013.

[5] M.-C. Yang and H.-C. Rim, "Identifying interesting Twitter contents using topical analysis," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4330–4336, 2014.

[6] J. Berger and K. L. Milkman, "Social transmission, emotion, and the virality of online content," Wharton Research Paper, 2010, http://www.msi.org/reports/social-transmission-emotion-and-the-virality-of-online-content/.

[7] F. Squazzoni, W. Jager, and B. Edmonds, "Social simulation in the social sciences: a brief overview," *Social Science Computer Review*, vol. 32, no. 3, pp. 279–294, 2014.

[8] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: conversational aspects of retweeting on twitter," in *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences (HICSS '10)*, pp. 51–60, January 2010.

[9] Z. Yang, J. Guo, K. Cai et al., "Understanding retweeting behaviors in social networks," in *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 1633–1636, October 2010.

[10] T. R. Zaman, R. Herbrich, J. V. Gael, and D. Stern, "Predicting information spreading in twitter," in *Proceedings of the Workshop on Computational Social Science and the Wisdom of Crowds*, pp. 35–42, 2010.

[11] D. Stern, R. Herbrich, and T. Graepel, "Matchbox: large scale online Bayesian recommendations," in *Proceedings of the 18th International World Wide Web Conference (WWW '09)*, pp. 111–120, April 2009.

[12] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network," in *Proceedings of IEEE 2nd International Conference on Social Computing (SocialCom '10)*, pp. 177–184, 2010.

[13] L. Zhang, Z. Zhang, and P. Jin, "Classification-based prediction on the retweet actions over microblog dataset," in *Proceedings of the 13th International Conference on Web Information Systems Engineering*, pp. 771–776, 2012.

[14] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to win! Predicting message propagation in twitter," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pp. 586–589, 2011.

[15] M. Cheong and V. Lee, "A study on detecting patterns in Twitter intra-topic user and message clustering," in *Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10)*, pp. 3125–3128, tur, August 2010.

[16] Y. He, W. Su, Y. Tian, Q. Chen, and L. Lin, "Summarizing microblogs on network hot topics," in *Proceedings of the International Conference on Internet Technology and Applications (iTAP '11)*, pp. 1–4, IEEE, Wuhan, China, August 2011.

[17] D. Zhang, Y. Liu, R. D. Lawrence, and V. Chenthamarakshan, "ALPOS: a machine learning approach for analyzing microblogging data," in *Proceedings of the 10th IEEE International Conference on Data Mining Workshops (ICDMW '10)*, pp. 1265–1272, December 2010.

[18] A. Celikyilmaz, D. Hakkani-Tür, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT '10)*, pp. 79–84, December 2010.

[19] Y. Wu and F. Ren, "Learning sentimental influence in twitter," in *Proceedings of the International Conference on Future Computer Sciences and Application*, pp. 119–122, June 2011.

[20] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang, "Social action tracking via noise to tolerant time-varying factor graphs," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 1049–1058, July 2010.

[21] Y. Artzi, P. Pantel, and M. Gamon, "Predicting responses to microblog posts," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 602–606, 2012.

[22] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM '12)*, pp. 26–33, June 2012.

[23] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in Twitter," in *Proceedings of the 20th International Conference Companion on World Wide Web*, pp. 57–58, April 2011.

[24] Y. Bae, P. Ryu, and H. Kim, "Predicting the lifespan and retweet times of tweets based on multiple feature analysis," *ETRI Journal*, vol. 36, no. 3, pp. 418–428, 2014.

[25] O. Tsur and A. Rappoport, "What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM '12)*, pp. 643–652, February 2012.

[26] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 46–50, August 2012.

[27] J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li, "Social influence locality for modeling retweeting behaviors," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, pp. 2761–2767, August 2013.

[28] B. Krishnamurthy, P. Gill, and M. Arlitt, "A few chirps about Twitter," in *Proceedings of the 1st Workshop on Online Social Networks (WOSP '08)*, pp. 19–24, August 2008.

[29] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar, "Detecting uninteresting content in text streams," in *Proceedings of the SIGIR Workshop on Crowdsourcing for Search Evaluation*, pp. 39–42, 2010.

[30] W. Webberley, S. Allen, and R. Whitaker, "Retweeting: a study of message-forwarding in twitter," in *Proceedings of the 1st NSS Workshop on Mobile and Online Social Networks (MOSN '11)*, pp. 13–18, September 2011.

[31] M. S. Tang, X. J. Mao, S. Q. Yang, and H. P. Zhou, "A dynamic microblog network and information dissemination in '@' mode," *Mathematical Problems in Engineering*, vol. 2014, Article ID 492753, 15 pages, 2014.

[32] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[33] M. Yavas and G. Yücel, "Impact of homophily on diffusion dynamics over social networks," *Social Science Computer Review*, vol. 32, no. 3, pp. 354–372, 2014.