Hindawi
Mathematical Problems in Engineering
Volume 2017, Article ID 1383891, 12 pages
https://doi.org/10.1155/2017/1383891

Research Article

SHMF: Interest Prediction Model with Social Hub Matrix Factorization

Chaoyuan Cui,[1] Hongze Wang,[2] Yun Wu,[3] Sen Gao,[4] and Shu Yan[1]

[1]*Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui 230031, China*
[2]*University of Chinese Academy of Sciences, Beijing 100049, China*
[3]*Institute of Applied Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui 230088, China*
[4]*University of Science and Technology of China, Hefei, Anhui 230031, China*

Correspondence should be addressed to Shu Yan; yanshu@iim.ac.cn

Received 24 January 2017; Accepted 5 June 2017; Published 22 August 2017

Academic Editor: Zonghua Zhang

With the development of social networks, microblog has become the major social communication tool. There is a lot of valuable information such as personal preference, public opinion, and marketing in microblog. Consequently, research on user interest prediction in microblog has a positive practical significance. In fact, how to extract information associated with user interest orientation from the constantly updated blog posts is not so easy. Existing prediction approaches based on probabilistic factor analysis use blog posts published by user to predict user interest. However, these methods are not very effective for the users who post less but browse more. In this paper, we propose a new prediction model, which is called SHMF, using social hub matrix factorization. SHMF constructs the interest prediction model by combining the information of blogs posts published by both user and direct neighbors in user's social hub. Our proposed model predicts user interest by integrating user's historical behavior and temporal factor as well as user's friendships, thus achieving accurate forecasts of user's future interests. The experimental results on Sina Weibo show the efficiency and effectiveness of our proposed model.

1. Introduction

Online microblog systems such as Sina Weibo, Twitter, and Facebook provide a convenient platform for users to share their information. The number of such social media users showed exponential growth in last decade. A recent snapshot of the friendship network Facebook indicated that there are over 1 billion users in it. These social networks are becoming not only effective means to connect their friends but also powerful information dissemination and marketing platforms to spread ideas, fads, and political opinions.

Microblog contains a vast amount of information, and topics of users and user groups always change with hotspot at home and abroad or over time. In this context, research on user interest prediction is useful in network marketing, public opinion analysis, or even public security [1]. Generally, interest prediction is to generate potential and possible topics in the next time point according to one's historical blog posts. Unfortunately, blog posts are almost short text; both user-keyword matrix and user-topic matrix of microblogs are relatively very sparse. Moreover, in the prediction model, contents of the related matrices transfer with lots of factors, such as time information and friendship in social hub. Therefore, interest prediction is still a challenging problem.

It should be noted that user interest prediction is different from user interest detection, as the latter mainly focuses on mining users' current interests. Interest prediction remains a relatively understudied problem that poses two main challenges. First, user interest in microblog changes over time or time interval. In the time-aware prediction model, user's temporal preference is an important aspect. Furthermore, long-term preference and short-term preference will result in different prediction result. Second, user interest is a dynamic phenomenon; it maybe migrates due to the topic migration of one's social hub. In the real world, capturing user's friendship and their topics is difficult.

Recently, a lot of models for prediction have been investigated [2–4]. A typical method exploits the probabilistic

matrix factorization (PMF) technique to learn latent features for users and topics. These kinds of algorithms are mostly based on the blog posts published by user to predict his interest.

In fact, we observed several interesting phenomena. There exist some users who publish less but browse more blog posts and we call them silent type users. Such users may have very explicit interest and just may be prudent to express their ideas. And they do publish their opinion at an appropriate moment. However, existing prediction models always fail to predict their interests. Another kind of users expands their social hubs by focusing on new friends' topics they are interested in. We call them interactive type users. In other words, the interest of such users can be represented by the interest of direct neighbors in their social hubs to some extent. Obviously, prediction models ignoring the impact of this interactive property always result in incomplete forecast.

In order to overcome the shortcomings of existing works, combining our observations about microblog, this paper proposes a social hub matrix factorization-based model for user interest prediction model in microblog, which is called SHMF. SHMF incorporates the impact of user's social hub on user's interests in our model to improve the quality of prediction. The experimental results on Sina Weibo dataset show that our approach improves the prediction accuracy and the performance efficiency.

The rest of this paper is organized as follows. The related work is discussed in Section 2. Some preliminary knowledge and research are introduced in Section 3. We present our proposed model in Section 4 and give the implementation details in Section 5. In Section 6, we describe the real datasets we used in our experiments. Our experiments are reported in Section 7. Finally, we conclude the paper and present some directions for future work in Section 8.

2. Related Work

With regard to user interest prediction in microblog, there are a series of mature methods that are based on probability matrix factorization of probabilistic graph model. Probabilistic graph model is a kind of model which can concisely express complex probability distribution, effectively calculate the edge and condition distribution, and conveniently learn the parameters and hyperparameters in probability model [5], while probability matrix factorization based on this model is often used to predict the user's interests and recommendations.

In 2008, Salakhutdinov and Mnih [2] proposed a probability matrix factorization (PMF) method for the traditional collaborative filtering algorithm which cannot solve the problem of the recommendation of large sparse dataset and cold start. Experiments on datasets of Netflix demonstrate the effectiveness of PMFs on large number of sparse unbalanced datasets. In the same year, Ma et al. [3] applied PMF to social network and socialization recommendation and analyzed the complexity and prediction accuracy of this method in detail. In 2010, combining the characteristics of social networks, Jamali and Ester [4] proposed a social probability matrix factorization (SocialMF) model based on the consideration of the social trust relationship between users. This model promotes the application prospect of PMF in socialization recommendation. In 2003, Sun et al. [6] proposed a method to model the user's timing behavior and combined this method with the SocialMF to predict the Weibo user's interest, the experimental results of which prove that this way of modeling is more effective than the traditional recommendation algorithm based on label information. Taking into account the fact that user interest is changing over time, Bao et al. [7] introduced a new temporal and social PMF-based (TS-PMF) method to predict users' interests in microblog. Compared with previous methods of interest prediction, this method has higher accuracy.

The above studies neglect the impact of the information of the blogs posted by others in their social hub on the user's future interest and behavior, when they establish the Weibo user interest prediction model. Aiming at this problem, in this paper, we propose a new user interest prediction model (SHMF) based on PMF, which combines user's history behavior, user's social trust relationship, and the impact of the information of the users' social hub on the user's interests in the future. And it designs experiments on the Sina microblog real dataset to prove that this prediction model and the algorithm of the model are superior to the previous prediction model in top-n accuracy [8].

3. Preliminaries

In this section, we give the notations that will be used in the following discussions. In prediction model, we have a set of users $\{u_1, u_2, \ldots, u_n\}$ and a set of topics $\{v_1, v_2, \ldots, v_m\}$ in a microblog dataset.

The users' interests expressed by user-topic matrix are given in $R \in R^{n*m}$, where $r_{ij} = 1$ if user u_i has published posts on topic v_j. We divide users' historical data into N time points $(T_1, T_2, \ldots, T_t)$ and construct a set of user-topic matrix $\mathbb{R}_1 = \{R_{11}, R_{12}, \ldots, R_{1t}\}$ to represent user's interests over time. Furthermore, considering the impact of user's social hub on his/her interest, we can construct a set of user's social hub-topic matrix $\mathbb{R}_2 = \{R_{21}, R_{22}, \ldots, R_{2t}\}$ according to the blogs posted by friends of his/her social hub.

In microblog, each user can follow others whom he is interested in; then users' friendships can be described as a user-user matrix $F_1 \in R^{n*n}$, where $F_{1,ij} = 1$ which denotes that u_i has followed u_j. Each user can mainly read the blogs posted by his friends of his social hub. Obviously, there are interactions among different users' social hubs. Users' social hubs can be described as a hub-hub matrix $F_2 \in R^{n*n}$. We set $F_{2,ij} = n_{ij}/n_i$ if the number of users in the intersection of hub α_i and hub α_j is n_{ij} and the number of users in hub α_i is n_i. Hub α_i is a set of users who are followed by u_i, and we have a set of user social hubs $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$.

Generally, user interest prediction model is to generate a user-interest matrix in the next time segment. The basic matrix factorization (MF) approach finds the approximate matrix of the original matrix in the low-rank space as a predictive approximation matrix. It has been proven to be effective to learn the latent characteristics of users and topics

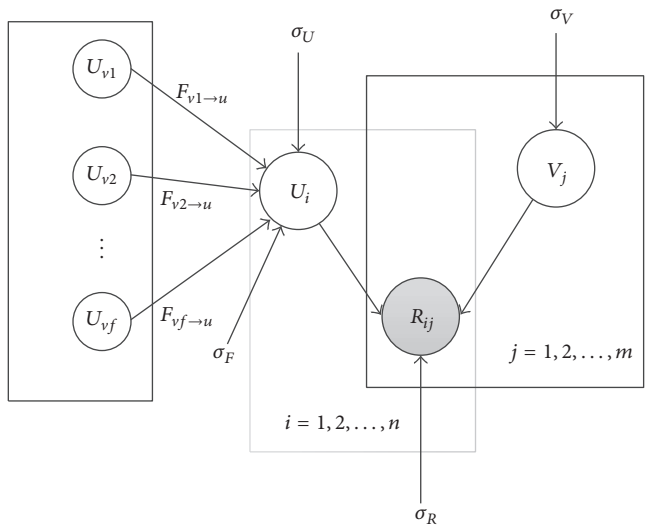

Figure 1: Graphical model of SocialMF.

and predict the scores using these latent characteristics. The conditional probability of the known scores is defined as

$$P\left(R \mid U, V, \sigma_R^2\right) = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[N\left(r_{ij} \mid g\left(U_i^T V_j\right), \sigma_R^2\right)\right]^{I_{ij}^R}. \quad (1)$$

As is shown in (1), $U \in R^{d*n}$ and $V \in R^{d*m}$ are the latent characteristics of users and topic feature matrices, with column vectors U_i and V_j representing d-dimensional user-latent and topic-latent feature vectors, respectively; $r_{ij} \approx U_i^T V_j$, where U_i^T is the transpose of U_i. $N(x \mid \mu, \sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2, and I_{ij}^R is the indicator function that is equal to 1 if $r_{ij} = 1$ and is equal to 0 otherwise. The function $g(x)$ is a logistic function with the formula $g(x) = 1/(1 + \exp(-x))$, which makes it possible to bound x within the range $[0, 1]$.

In fact, the relations among users in social network architecture play an important role in users' behaviors [9, 10]. Specifically, a user is more and more similar to his/her friends. SocialMF model incorporates social influence into the MF approach for prediction, adding the user-user relationship matrix $F \in R^{n*n}$:

$$\begin{aligned} P\left(U \mid F, \sigma_U^2, \sigma_F^2\right) &\propto P\left(U \mid \sigma_U^2\right) * P\left(U \mid F, \sigma_F^2\right) \\ &= \prod_{i=1}^{n} N\left(U_i \mid 0, \sigma_U^2 I\right) \\ &\quad * \prod_{i=1}^{n} N\left(U_i \mid \sum_j F_{ij} U_j, \sigma_F^2 I\right). \end{aligned} \quad (2)$$

Figure 1 shows the graphical model corresponding to (2). In Figure 1, the edges among the latent feature vectors of users are representatives of the trust relationship among users and the degree of trust of user u on user v is $F_{u\to v}$.

The user-topic matrices in PMF and SocialMF model are all constructed from the user's historical behavior information and do not take time influence into account. Meanwhile TS-PMF model incorporates characteristics of the user interest over time and adds the exponential decay function to analyze the user-topic matrices [7]. TS-PMF is designed to utilize users' sequential interest matrices $\{R_{11}, R_{12}, \ldots, R_{1t}\}$ and the users' friendships matrix $F \in R^{n*n}$ to predict users' interest in the near future. In time t, the conditional distribution probability of the observed items in R_{1t} is similar to that in (1):

$$\begin{aligned} &P\left(R_t \mid U_t, V_t, \sigma_{R_t}^2\right) \\ &\quad = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[N\left(R_{tij} \mid g\left(U_{t,j}^T V_{t,j}\right), \sigma_{R_t}^2\right)\right]^{I_{ij}^{R_t}}. \end{aligned} \quad (3)$$

Adding the exponential decay function to analyze the change of user interest, the computing formulation is listed as follows:

$$\begin{aligned} M_{U_t} &= \theta \sum_{k=1}^{t-1} \exp\left(\frac{t-k}{\beta}\right) U_k, \\ M_{V_t} &= \theta \sum_{k=1}^{t-1} \exp\left(\frac{t-k}{\beta}\right) V_k. \end{aligned} \quad (4)$$

The user's latent feature vector is affected by his historical interests and his friends' interests. Therefore, the conditional distribution probability of users' latent features can be expressed like this:

$$\begin{aligned} &P\left(U_t \mid \{R_1, R_2, \ldots, R_{t-1}\}, F, \sigma_{U_t}^2, \sigma_{F_t}^2\right) \\ &\quad \propto P\left(U_t \mid \{R_1, R_2, \ldots, R_{t-1}\}, \sigma_{U_t}^2\right) * P\left(U_t \mid F, \sigma_F^2\right) \\ &\quad = \prod_{i=1}^{n} N\left(U_{t,i} \mid M_{U_{i,j}}, \sigma_{U_t}^2 I\right) \\ &\qquad * \prod_{i=1}^{n} N\left[U_i \mid \sum_j F_{ij} U_j, \sigma_F^2 I\right]. \end{aligned} \quad (5)$$

Now, through a Bayesian inference, we have the following equation for the posterior probability over latent features of users and topics:

$$\begin{aligned} &P\left(U_t, V_t \mid \{R_1, R_2, \ldots, R_t\}, F, \sigma_{U_t}^2, \sigma_{V_t}^2, \sigma_F^2, \sigma_{R_t}^2\right) \\ &\quad \propto P\left(R_t \mid U_t, V_t, \sigma_{R_t}^2\right) \\ &\qquad * P\left(U_t \mid \{R_1, R_2, \ldots, R_{t-1}\}, \sigma_{U_t}^2\right) \\ &\qquad * P\left(U_t \mid F, \sigma_F^2\right) \\ &\qquad * P\left(V_t \mid \{R_1, R_2, \ldots, R_{t-1}\}, \sigma_{V_t}^2\right). \end{aligned} \tag{6}$$

Maximizing the log of the posterior distribution with regard to U_t and V_t is equivalent to minimizing the following sum-of-squared-errors objective function (we can find a local optimal value of the objective function by performing gradient descent):

$$\begin{aligned} &E\left(U_t, V_t \mid \{R_1, R_2, \ldots, R_t\}, F\right) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}^{R_t}\left(R_{t,ij} - g\left(U_{t,j}^T V_{t,j}\right)\right)^2 \\ &\quad + \frac{\lambda_{U_t}}{2}\left\|U_t - M_{U_t}\right\|_F^2 + \frac{\lambda_{V_t}}{2}\left\|U_t - M_{V_t}\right\|_F^2 \\ &\quad + \frac{\lambda_F}{2}\sum_{i=1}^{n}\left(U_{t,i} - \sum_v F_{iv}U_{t,v}\right)^T\left(U_{t,i} - \sum_v F_{iv}U_{t,v}\right). \end{aligned} \tag{7}$$

4. Social Hub User Interest Prediction Model

In this section, we present our model, SHMF, to incorporate impact of user's social hub into MF approach for prediction. SHMF combines user's historical behavior, social trust relationship, and blog articles posted by friends in user's social hub.

Independence Hypothesis. Information of blogs posted in users' social hub influences users' interests independently.

Based on the above hypothesis, we have

$$\begin{aligned} &P\left(R_{1t}, R_{2t} \mid U_t, V_t, \sigma_{R_{1t}}^2, \sigma_{R_{2t}}^2\right) \\ &\quad \propto P\left(R_{1t} \mid U_t, V_t, \sigma_{R_{1t}}^2\right) * P\left(R_{2t} \mid U_t, V_t, \sigma_{R_{2t}}^2\right) \\ &\quad = \prod_{i=1}^{n}\prod_{j=1}^{m}\left[N\left(R_{1t,ij} \mid g\left(U_{t,i}^T V_{t,j}\right), \sigma_{R_{1t}}^2\right)\right]^{I_{ij}^{R^{1t}}} \\ &\qquad * \prod_{i=1}^{n}\prod_{j=1}^{m}\left[N\left(R_{2t,ij} \mid g\left(U_{t,i}^T V_{t,j}\right), \sigma_{R_{2t}}^2\right)\right]^{I_{ij}^{R^{2t}}}. \end{aligned} \tag{8}$$

Therefore, the conditional distribution probability of users' latent features can be expressed as follows:

$$\begin{aligned} &P\left(U_t \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_1, F_2, \right. \\ &\left.\sigma_{U_{1t}}^2, \sigma_{U_{2t}}^2, \sigma_{F_1}^2, \sigma_{F_2}^2\right) \propto P\left(U_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, F_1, \right. \\ &\left.\sigma_{U_{1t}}^2, \sigma_{F_1}^2\right) * P\left(U_t \mid \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_2, \sigma_{U_{2t}}^2, \sigma_{F_2}^2\right). \end{aligned} \tag{9}$$

Through a Bayesian inference, we have the following equation for the posterior probability over latent features of users and topics:

$$\begin{aligned} &P\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_1, \right. \\ &\left. F_2, \sigma_{U_{1t}}^2, \sigma_{U_{2t}}^2, \sigma_{V_{1t}}^2, \sigma_{V_{2t}}^2, \sigma_{F_1}^2, \sigma_{F_2}^2\right) = P\left(U_t, V_t \mid \right. \\ &\left.\{R_{11}, R_{12}, \ldots, R_{1(t)}\}, F_1, \sigma_{U_{1t}}^2, \sigma_{V_{1t}}^2, \sigma_{F_1}^2\right) * P\left(U_t, V_t \mid \right. \\ &\left.\{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_2, \sigma_{U_{2t}}^2, \sigma_{V_{2t}}^2, \sigma_{F_2}^2\right). \end{aligned} \tag{10}$$

The log of the posterior distribution for SHMF at time point t is given by

$$\begin{aligned} &\ln\left(P\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, \right. \right. \\ &\left.\left. F_1, F_2, \sigma_{U_{1t}}^2, \sigma_{U_{2t}}^2, \sigma_{V_{1t}}^2, \sigma_{V_{2t}}^2, \sigma_{F_1}^2, \sigma_{F_2}^2\right)\right) = \ln\left(P\left(U_t, V_t \mid \right. \right. \\ &\left.\left.\{R_{11}, R_{12}, \ldots, R_{1(t)}\}, F_1, \sigma_{U_{1t}}^2, \sigma_{V_{1t}}^2, \sigma_{F_1}^2\right)\right) + \ln\left(P\left(U_t, \right. \right. \\ &\left.\left. V_t \mid \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_2, \sigma_{U_{2t}}^2, \sigma_{V_{2t}}^2, \sigma_{F_2}^2\right)\right). \end{aligned} \tag{11}$$

Maximizing the log of the posterior distribution with regard to U_t and V_t is equivalent to minimizing the following sum-of-squared-errors objective function:

$$\begin{aligned} &E\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_1, \right. \\ &\left. F_2\right) = E_1\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, F_1\right) + E_2\left(U_t, \right. \\ &\left. V_t \mid \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_2\right). \end{aligned} \tag{12}$$

In (12), E_1 and E_2 can be computed by (7). It is obvious that SHMF interest prediction is actually equivalent to performing the symmetrical calculation on the loss function. Here we introduce a parameter $\lambda \in [0, 1]$ to indicate the importance of user's social hub information in user's interest. We set $\lambda = 0$ if only user's personal posting behavior is considered and set $\lambda = 1$ if only user's social hub information is considered. Thus, the loss function can be computed as follows:

$$\begin{aligned} &E\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_1, \right. \\ &\left. F_2, \lambda\right) = (1 - \lambda) * E_1\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1(t)}\}, F_1\right) \\ &\quad + \lambda * E_2\left(U_t, V_t \mid \{R_{21}, R_{22}, \ldots, R_{2(t)}\}, F_2\right), \\ &E_1\left(U_t, V_t \mid \{R_{11}, R_{12}, \ldots, R_{1t}\}, F_1\right) = \frac{1}{2} \\ &\quad \cdot \sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}^{R_{1t}}\left(R_{1t,ij} - g\left(U_{t,i}^T V_{t,j}\right)\right)^2 + \frac{\lambda_{U_{1t}}}{2}\left\|U_t \right. \\ &\quad \left. - M_{U_t}\right\|_F^2 + \frac{\lambda_{V_{1t}}}{2}\left\|V_t - M_{V_t}\right\|_F^2 + \frac{\lambda_{F_1}}{2} \\ &\quad \cdot \sum_{i=1}^{n}\left(U_{t,i} - \sum_v F_{1iv}U_{t,v}\right)^T\left(U_{t,i} - \sum_v F_{1iv}U_{t,\mathrm{v}}\right), \end{aligned}$$

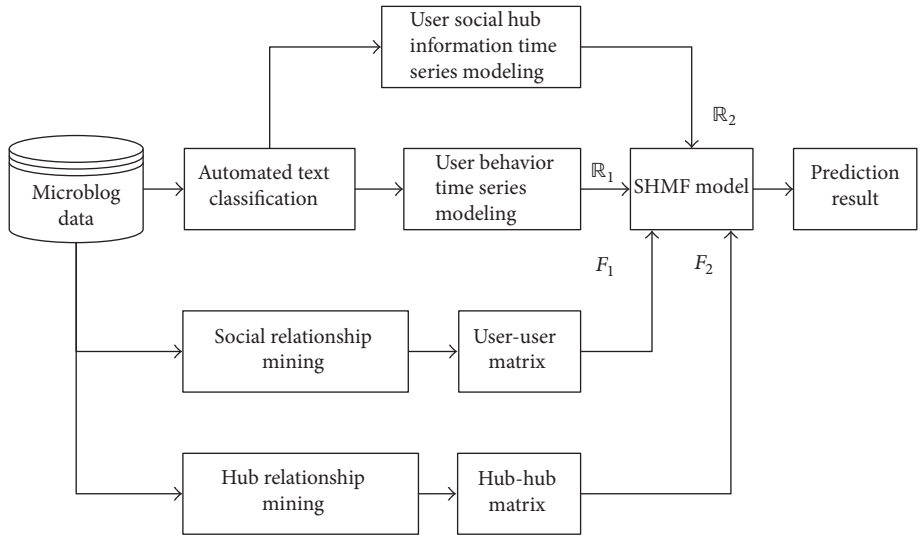

Figure 2: The framework of predicting users' interests.

$$E_2\left(U_t, V_t \mid \{R_{21}, R_{22}, \ldots, R_{2t}\}, F_2\right) = \frac{1}{2} \cdot \sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}^{R_{2t}}\left(R_{2t,ij} - g\left(U_{t,i}^T V_{t,j}\right)\right)^2 + \frac{\lambda_{U_{2t}}}{2}\left\|U_t - M_{U_t}\right\|_F^2 + \frac{\lambda_{V_{2t}}}{2}\left\|V_t - M_{V_t}\right\|_F^2 + \frac{\lambda_{F_2}}{2} \cdot \sum_{i=1}^{n}\left(U_{t,i} - \sum_v F_{2iv}U_{t,v}\right)^T\left(U_{t,i} - \sum_v F_{2iv}U_{t,v}\right). \tag{13}$$

In order to reduce the computational complexity, stochastic gradient descent is used to optimize the local optimum of the loss function, as shown in (14):

$$U_t \coloneqq U_t + \alpha\left((1-\lambda)U\delta_1 + \lambda U\delta_2\right), \quad V_t \coloneqq U_t + \alpha\left((1-\lambda)V\delta_1 + \lambda V\delta_2\right), \tag{14}$$

$$U\delta_1 = I_{ij}^{R_{1t}} g'\left(U_{t,i}^T V_{t,j}\right)\left(R_{1t,ij} - g\left(U_{t,i}^T V_{t,j}\right)\right)V_{t,j} - \lambda_{U_{1t}}\left(U_{t,i} - M_{U_{t,i}}\right) - \lambda_{F_1}\left(1 - F_{1,ii}\right)\left(U_{t,i} - \sum_v F_{1,iv}U_{t,v}\right), \tag{15}$$

$$U\delta_2 = I_{ij}^{R_{2t}} g'\left(U_{t,i}^T V_{t,j}\right)\left(R_{2t,ij} - g\left(U_{t,i}^T V_{t,j}\right)\right)V_{t,j} - \lambda_{U_{2t}}\left(U_{t,i} - M_{U_{t,i}}\right) - \lambda_{F_2}\left(1 - F_{2,ii}\right)\left(U_{t,i} - \sum_v F_{2,iv}U_{t,v}\right), \tag{16}$$

$$V\delta_1 = I_{ij}^{R_{1t}} g'\left(U_{t,i}^T V_{t,j}\right)\left(R_{1t,ij} - g\left(U_{t,i}^T V_{t,j}\right)\right)U_{t,j} - \lambda_{V_{1t}}\left(V_{t,i} - M_{U_{t,i}}\right), \tag{17}$$

$$V\delta_2 = I_{ij}^{R_{2t}} g'\left(U_{t,i}^T V_{t,j}\right)\left(R_{2t,ij} - g\left(U_{t,i}^T V_{t,j}\right)\right)U_{t,j} - \lambda_{V_{2t}}\left(V_{t,i} - M_{U_{t,i}}\right), \tag{18}$$

where $g'(x) = \exp(-x)/(1+\exp(-x))^2$ is the first-order derivative of logistic function $g(x)$; $\lambda_{F_a} = \sigma_{R_{at}}^2/\sigma_{F_a}^2$, $\lambda_{U_{at}} = \sigma_{R_{at}}^2/\sigma_{U_{at}}^2$, $\lambda_{V_{at}} = \sigma_{R_{at}}^2/\sigma_{V_{at}}^2$, $\alpha = 1, 2$, and $\|\cdot\|_F^2$ are the Frobenius norm.

SHMF model provides an effective way to predict users' interests. The procedure of prediction will be described with two algorithms in Section 5. All the notations used throughout the paper are summarized in Notations.

5. Implementation

To evaluate the effectiveness and efficiency of our approach, we implemented a prototype system of user interest prediction. According to SHMF model and its variant, we provide two algorithms with different parameters and procedures.

5.1. Architecture Overview. The architecture of our implementation is illustrated in Figure 2. We first use topic model LDA to mark out topics of the microblog dataset automatically. Meanwhile, we use the sequential behaviors of users to get a set of user-topic matrices $\mathbb{R}_1 = \{R_{11}, R_{12}, \ldots, R_{1t}\}$ and a set of users' social hub-topic matrices $\mathbb{R}_2 = \{R_{21}, R_{22}, \ldots, R_{2t}\}$. Next, we capture the social relationship between users and get a user-user matrix $F_i \in R_{n*n}$ and we can get a hub-hub matrix in the same way. Finally, $\mathbb{R}_1, \mathbb{R}_2, F_1$, and F_2 are input to the SHMF model to generate the prediction result.

5.2. Algorithms. SHMF integrates user's history behavior, user's social trust relationship, and the impact of the information of user's social hub. The process of predicting users' interests with SHMF is described in Algorithm 1.

Require:
Dataset: $\{R_{11}, R_{12}, \ldots, R_{1N}\}, \{R_{21}, R_{22}, \ldots, R_{2N}\} F_1, F_2$;
The dimension of the latent feature: d;;
Parameters: $\lambda_{U_1}, \lambda_{V_1}, \lambda_{F_1}, \lambda_{U_2}, \lambda_{V_2}, \lambda_{F_2}, \theta, \beta, \lambda$;
An updating parameter: α
Convergence parameter: ε
The maximum number of iterations: K
Ensure:
The user-topic matrix in time segment $N + 1$: R_{N+1}
(1) $M_{1U_1} = \text{zeros}(d, n)$, $M_{1V_1} = \text{zeros}(d, m)$, $M_{2U_1} = \text{zeros}(d, n)$, $M_{2V_1} = \text{zeros}(d, m)$
(2) **for** $t = 1, \ldots, N$ **do**
(3) initialize $U_t, V_t : U_{t,0} = U_t$, $V_{t,0} = V_t$, $E_0 = \text{inf}$;
(4) **if** $t > 1$ **then**
(5) Compute the mean matrices M_{1U_t}, M_{1V_t}, M_{2U_t}, M_{2V_t}
(6) **end if**
(7) **for** $l = 1, \ldots, K$ **do**
(8) compute the gradient descent in Eq. (15) (16) (17) (18);
(9) updating in Eq. (14);
(10) compute E in Eq. (13);
(11) **if** $|E_0 - E| < \varepsilon$ **then**
(12) break
(13) **end if**
(14) **if** $E = \min\{E_0, E\}$ **then**
(15) $U_{t,0} = U_t$, $V_{t,0} = V_t$
(16) **end if**
(17) **end for**
(18) $U_t = U_{t,0}, V_t = V_{t,0}$
(19) **end for**
(20) predict R_{N+1} using $R_{N+1} \approx U_N^T V_N$

Algorithm 1: The process of predicting users' interests.

6. Datasets

6.1. Experimental Data. We used the dataset from 1 May 2016 to 31 May 2016, which we downloaded from Sina Weibo. This dataset includes more than 20 million microblog messages, time-stamps, and user-to-user relationships.

6.2. User Selection. The basic idea of traditional collaborative filtering is that similar users make similar choices, or similar options are chosen by similar groups of users [11]. In recent years, the basic idea of the social recommendations is gradually concerned by the researchers. The researchers of the social recommendations think that, for a social impact of consideration [12, 13], the associated users will affect each other, so the user's interest is largely influenced by the users associated with him.

Taking into account the complexity of the calculation, the selection of users is very important in the microblog user interest prediction. In a month, different users will post different numbers of microblogs. Someone only posts one, but someone posts tens of thousands. For such users who post little of microblog in a month, personal microblog information and social hub microblog information are unable to describe their interests. However, for the users who post lots of microblogs in a month, they mostly are enterprises and institutions of the official microblog or commercial procurement service, and it is meaningless to predict user's interest based on those users. To do this, we perform a statistical analysis on the dataset from Sina Weibo and find that the number of microblogs posted by most users is 100 or less as shown in Figures 3(a) and 3(b) showing histograms of the number of users with the different numbers of blog posts. In this paper, we select users who post 20 to 100 microblogs as subjects, and the number of this kind of users is about one million. After using neighbor computing [7] and stratified sampling, the 1402 users' information is selected as the experimental object.

6.3. Automatically Classify Blogs' Topics Posted by Users. After getting the user's blog information, we train the LDA model and use it to automatically classify the blogs posted by users and the blogs posted by others in user's social hub, and the number of topics is calculated by the perplexity. According to perplexity-numbers of topics curve shown in Figure 4, the best number of topics is 23 when the perplexity reached its lowest point.

7. Experiments and Analysis

In this section, effectiveness and efficiency of our SHMF model are evaluated. We conduct experiments on Intel Core i7 processor with 4 cores running at frequency of 3.60 GHz,

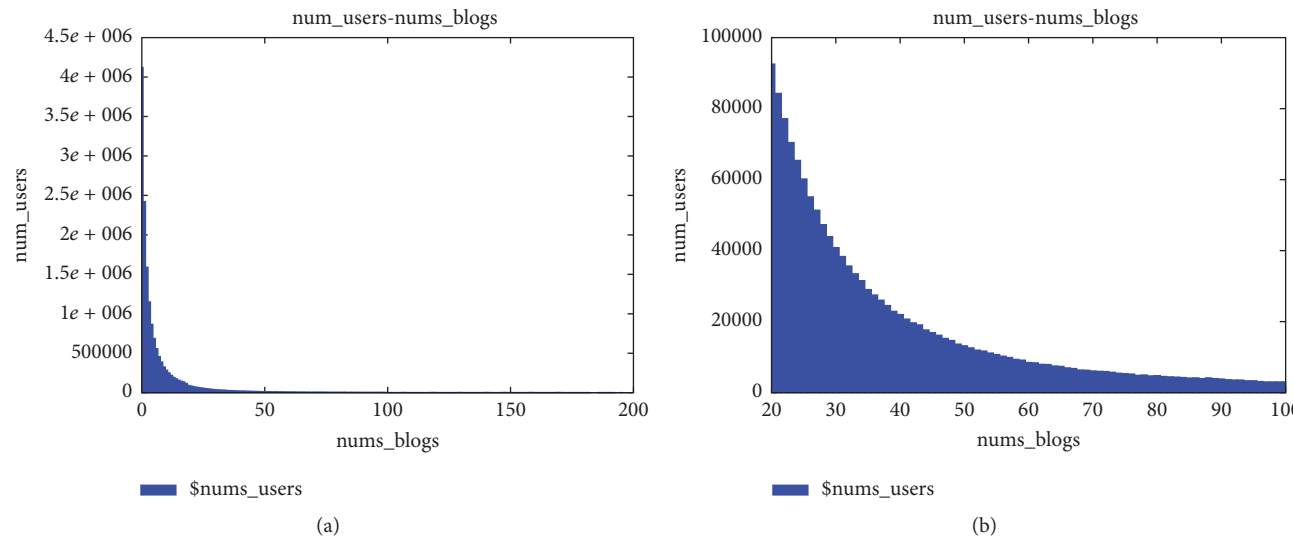

Figure 3: Statistical analysis of the dataset from Sina Weibo.

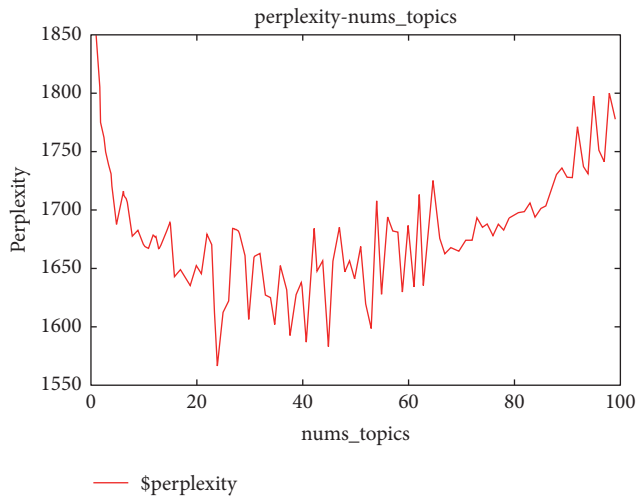

Figure 4: Perplexity-numbers of topics curve.

24 GB memory, and 1TB hard disk. The programs are run on Windows 7 Professional and Anaconda 4.1.1 (64-bit).

We first present evaluation metrics used throughout our experiments. Next, we employ the variable-controlling approach to adjust the parameters of SHMF model and the other three models. Then the prediction accuracy and the performance overhead of our model are compared with results of the other models. Finally, we will analyze the experimental results.

7.1. Metrics. Because of the great uncertainty of the behavior of user posting blogs, the recall rate has little practical significance in this issue, and in the real life users pay more attention to the top-N topic which they are most interested in. Therefore, in this paper, the precision of top-n is used as the model evaluation criteria:

$$\mathrm{Pre}_n = \frac{N_{\mathrm{correct}}(n)}{N_u \times n}. \tag{19}$$

N_u represents the number of users in the test set; and $N_{\mathrm{correct}}(n)$ represents the total number of interest topics predicted correctly in the top-n prediction results for all users in the corresponding test set.

7.2. Model Selection and Parameter Setting. We set up three experiments, PMF [2], SocialMF [4], and TS-PMF [7], as the contrastive experiments because these three methods are very often used to predict users' interests, and the three methods are in the same theoretical system as the model SHMF proposed in this paper. And then we set up an experiment for the model SHMF proposed in this paper.

First, the variable-controlling approach was used to adjust the parameters to better values, and then we compare their top-n accuracy and average accuracy.

(1) PMF Model. The PMF model has three parameters, λ_U, λ_V, d, in this paper; λ_U, λ_V are the regularization term coefficients in the loss function. The default value of λ_U, λ_V is 0.01 before setting parameters; d is the dimension of the latent features which is generally less than the rank of the original matrix. The control variable method is used to set the parameters by fixing other values and changing one. Then we can draw a graph to get the impact of each parameter on top-n accuracy. In order to reduce the computational complexity, we set $\lambda_U = \lambda_V$. The top-n accuracy varies with the parameters λ_U, λ_V, d as shown in Figure 5.

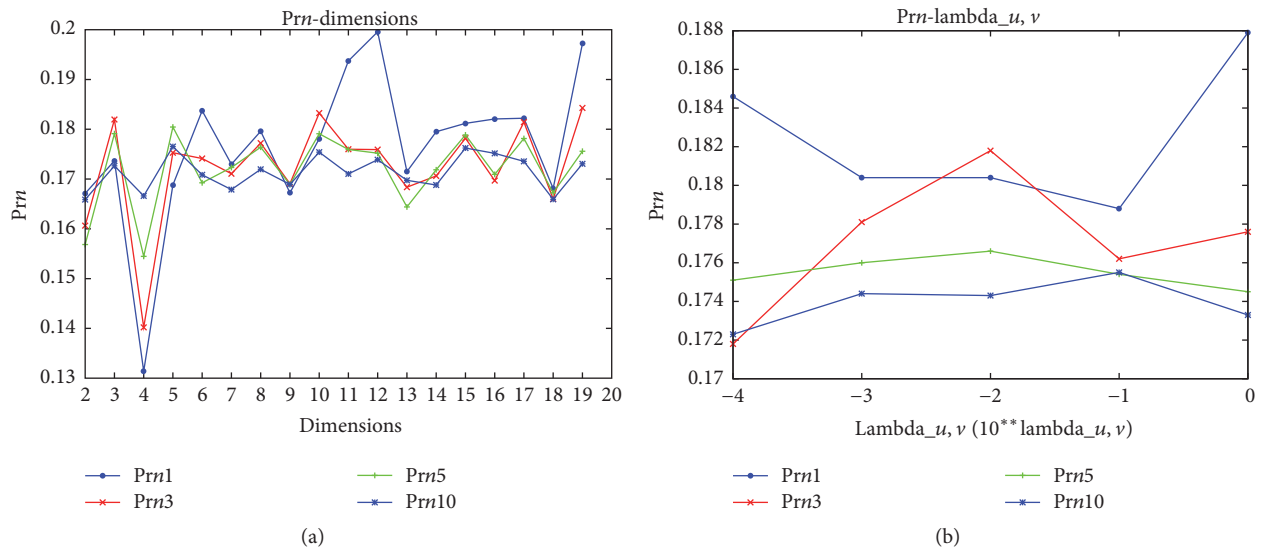

Figure 5: Impact of different values of different parameters in the PMF model on performance of user interest prediction.

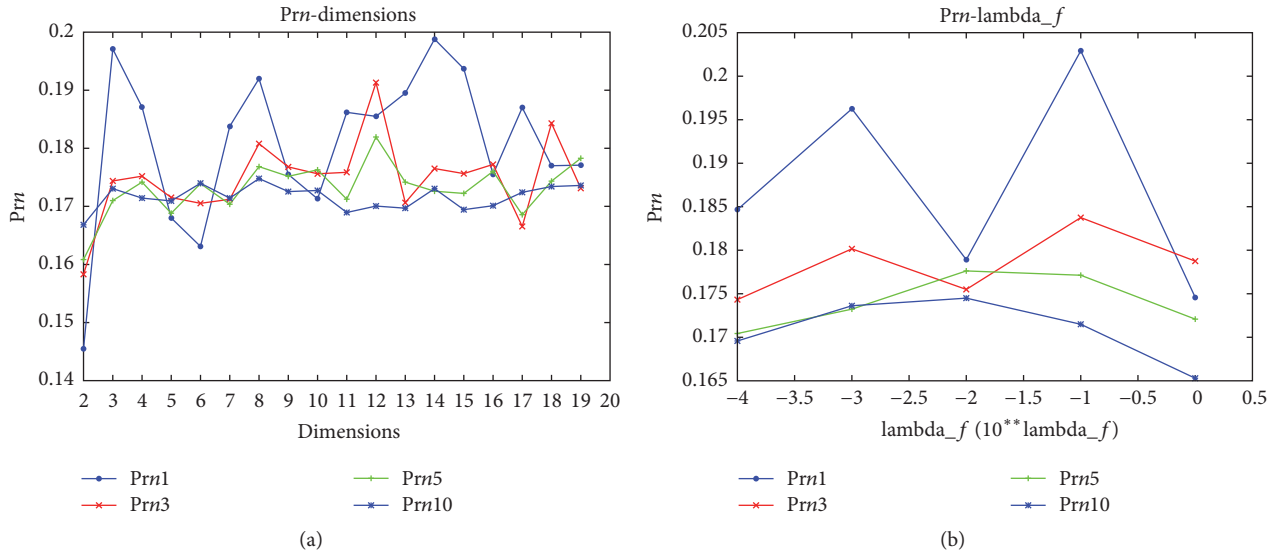

Figure 6: Impact of different values of different parameters in the SocialMF model on performance of user interest prediction.

According to Figure 5, we can get a set of parameters $d = 12$ and $\lambda_U = \lambda_V = 0.01$, which can make the models perform better on the top-1, top-3, top-5, and top-10 accuracy rate.

(2) SocialMF Model. The SocialMF model has four parameters, $\lambda_U, \lambda_V, \lambda_F, d$, in this paper. $\lambda_U, \lambda_V, \lambda_F$ are the regularization term coefficients in the loss function. In order to reduce the computational complexity, we set $\lambda_U = \lambda_V = 0.01$ which we set in the first experiment, and we set $\lambda_F = 0.001$ before setting parameters. d is the dimension of the latent features which is generally less than the rank of the original matrix. The control variable method is used to set the parameters by fixing other values and changing one. Then we can draw a graph to get the impact of each parameter on top-n accuracy. The top-n accuracy varies with the parameter d as shown in Figure 6(a) and with the parameter λ_F as shown in Figure 6(b).

Based on Figure 6, we can get a set of parameters $d = 12$, $\lambda_U = \lambda_V = 0.01$, and $\lambda_F = 0.1$ which can make the model have better performance on the top-1, top-3, top-5, and top-10 accuracy rate.

(3) TS-PMF Model. The TS-PMF model has six parameters, $\lambda_U, \lambda_V, \lambda_F, d, \theta$, and β. λ_U, λ_V, and λ_F are the regularization term coefficients in the loss function. In order to reduce the computational complexity, we set $\lambda_U = \lambda_V = 0.01$ which we set in the first experiment, and we set $\beta = 3$ and $\lambda = 0.001$ before setting parameters. d is the dimension of the latent features which is generally less than the rank of the original matrix. θ, β are the parameters in the forgotten function. The

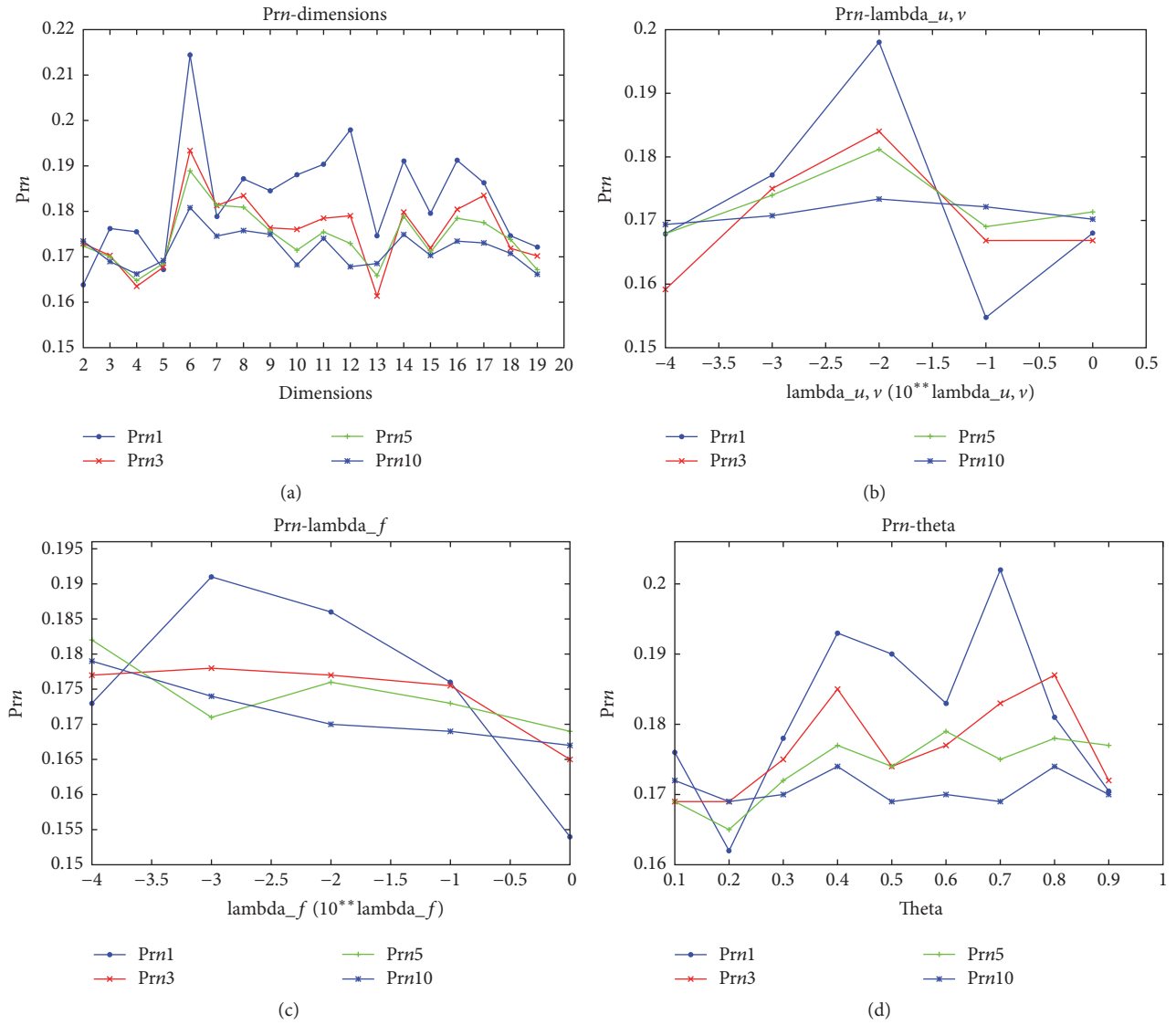

Figure 7: Impact of different values of different parameters in the TS-PMF model on performance of user interest prediction.

control variable method is used to set the parameters by fixing other values and changing one. Then we can draw a graph to get the impact of each parameter on top-n accuracy. The top-n accuracy varies with the parameters λ_U, λ_V, λ_F, d, and θ as shown in Figure 7.

From Figure 7, we can get a set of parameters $\lambda_U = \lambda_V = 0.01$, $\lambda_F = 0.001$, $d = 6$, and $\theta = 0.4$ which can make the model have better performance on the top-1, top-3, top-5, and top-10 accuracy rate.

(4) SHMF Model. The SHMF model has ten parameters, λ_{U_1}, λ_{V_1}, λ_{F_1}, λ_{U_2}, λ_{V_2}, λ_{F_2}, d, θ, β, and λ. In order to reduce the computational complexity, according to independence hypothesis, "the information of blogs posted by users and the information of blogs posted by others in user's social hub influence the user's interest in the future independently"; we can set $\lambda_{U_1} = \lambda_{V_1} = 0.01$ and $\lambda_{F_1} = 0.001$ in accordance with the third experiment. Then we set $\beta = 3$ and $\lambda_{U_2} = \lambda_{V_2}$, so we should actually consider the five parameters $\lambda_{U_2} = \lambda_{V_2}$, λ_{F_2}, d, θ, and λ, in which λ_{U_2}, λ_{V_2}, and λ_{F_2} are the regularization term coefficients in the loss function. d is the dimension of the latent features which is generally less than the rank of the original matrix. θ, β are the parameters in the forgotten function. λ indicates how important the user's social hub information is to the user's interest. We set $\lambda = 0$ if only user's personal posting behavior is considered and the SHMF model degrades to TS-PMF model at this time, and we set $\lambda = 1$ if only user's social hub information is considered. The control variable method is used to set the parameters by fixing other values and changing one. Then we can draw a graph to get the impact of each parameter on top-n accuracy. The top-n accuracy varies with the parameters λ_{U_2}, λ_{V_2}, λ_{F_2}, d, θ, and λ as shown in Figure 8.

According to Figure 8, we can get a set of parameters $\lambda_{U_1} = \lambda_{V_1} = 0.1$, $\lambda_{F_1} = 0.0001$, $d = 6$, $\theta = 0.3$, and $\lambda = 0.5$

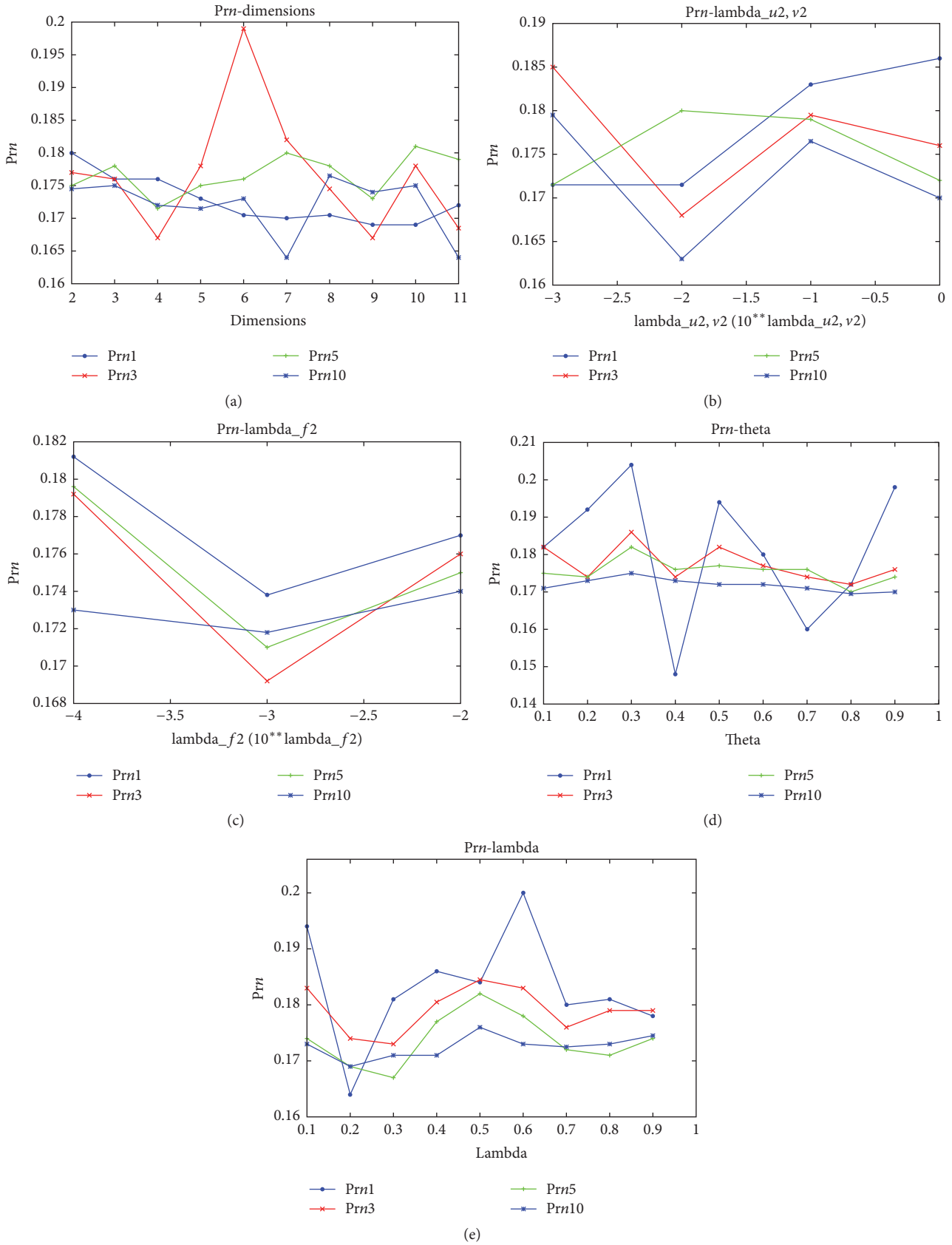

Figure 8: Impact of different values of different parameters in the SHMF model on performance of user interest prediction.

Table 1: Precision of SHMF.

	Pre_avg
PMF	17.35%
SocialMF	17.37%
TS-PMF	17.91%
SHMF	18.67%

Table 2: Performance of SHMF.

	Run-time (s)
PMF	698.618
SocialMF	1227.088
TS-PMF	1721.555
SHMF	2080.513

which can make the model have better performance on the top-1, top-3, top-5, and top-10 accuracy rate.

7.3. Experimental Results and Analysis

(1) Comparison of Accuracy. After adjusting the parameters of the five models, it is necessary to compare the strengths and weaknesses of the different models. As a result of the fact that the selection of different top-n accuracy will lead to different results, in order to consider comprehensively, this paper takes top-1, top-3, top-5, and top-10 accuracy of the arithmetic mean as the average accuracy, as shown in the following equation:

$$\mathrm{Pre}_{\mathrm{avg}} = \frac{\mathrm{Pre}_1 + \mathrm{Pre}_3 + \mathrm{Pre}_5 + \mathrm{Pre}_{10}}{4}. \tag{20}$$

By adjusting the model parameters of five experiments, the average accuracy of the five models under most parameters is shown in Table 1.

It can be seen from Table 1 that the algorithm SHMF proposed in this paper improves the average accuracy by over 1.3% compared to algorithm PMF and algorithm SocialMF and the average accuracy of the algorithm SHMF is 0.76% higher than the algorithm TS-PMF.

(2) Executive Efficiency Analysis. On the efficiency of implementation, based on the best parameters, set the number of iterations to 100 times and record the run-time, as shown in Table 2.

It is found from Table 2 that the running time of the algorithm SHMF is the longest, which is nearly three times the running time of the algorithm PMF. This is because, with the calculation of the complexity of the increase, the run-time of the algorithm SHMF has increased.

(3) Result Analysis. Through the comparison of four groups of experiments, we can see the difference and relation of PMF-based algorithm in microblog users' interest prediction. In the first comparative experiment, we use the most basic probability matrix factorization algorithm and got the average accuracy of 17.35%. In the second comparative experiment, the social trust relationship is added based on the probability matrix factorization algorithm. However, the average accuracy is almost the same as that obtained by the basic probability matrix factorization algorithm. This is mainly due to the fact that, in constructing dataset, we take the users whose posts are in a certain range and then determine their social trust relationships according to the statistical characteristics instead of using all or as many social trust relationships as possible for a user in order to consider both the similarity of behavior and the mutual influence among users. Therefore, this kind of method leads to sparsity of social trust matrix, so the impact is relatively small. Since we do not only focus on the correlation between users, we use this approach to implement the experiment. Compared with the previous two experiments, the average accuracy of the third comparative experiment is higher than that of the previous two experiments, and it is proven that the fact that this method based on the short-term interest of users is changing along time is rational. In the last experiment, the algorithm SHMF proposed in this paper will improve the average accuracy rate of nearly one percentage point, indicating that the user's social hub information does affect the user's interest in microblog and verifying the effectiveness of the algorithm at the same time.

8. Conclusions and Future Work

Based on the work of the prediction of microblog users' interest, this paper analyzes the information of microblog users' social hub and puts forward the SHMF model, which greatly improves the top-n accuracy and average accuracy. This will lay the foundation for the follow-up research work. At the same time, we can solve the cold-start problem of predicting interests of the users who do not often post blogs by analyzing the information of their social hub. This method could have a broad application space in social platform recommendation. However, there are still some defects in the implementation efficiency. When the amount of data is particularly large, the running time is too long, which needs to be improved in the future work.

For the future work of microblog users' interest prediction, further research on the expression of interest should be carried out to achieve more accurate representation, which determines the upper limit of interest prediction. In the prediction algorithm, we should add more techniques, such as Bayesian analysis, to solve the multiparameter problem by analyzing the relationship between the parameters and the actual meaning.

Notations

R_{1t}: The user-topic matrix in time t
R_{2t}: The user's social hub-topic matrix in time t
F_1: The user-user matrix
F_2: The hub-hub matrix
U_{1t}^T: The users' latent feature space in time t
V_{1t}^T: The topics' latent feature space in time t
U_{2t}^T: The users' latent feature space in social hub in time t

V_{2t}^T: The topics' latent feature space in social hub in time t
U_t^T: The final users' latent feature space in time t
V_t^T: The final topics' latent feature space in time t
$M_{U_{1t}}$: The mean matrix of U_{1t} with spherical Gaussian priors in time t
$M_{U_{2t}}$: The mean matrix of U_{2t} with spherical Gaussian priors in time t
$M_{V_{1t}}$: The mean matrix of V_{1t} with spherical Gaussian priors in time t
$M_{V_{2t}}$: The mean matrix of V_{2t} with spherical Gaussian priors in time t
θ: A weight that indicates how important the whole previous time points are to the current one
β: The kernel parameter
d: The dimension of latent feature space
λ: A weight that indicates how important the user's social hub information is to the user's interest
λ_{U_1}: The impact of the users' latent feature vectors on users' interests
λ_{U_2}: The impact of the social hubs' latent feature vectors on users' interests
λ_{V_1}: The impact of the topics of the blogs posted by users on users' interests
λ_{V_2}: The impact of the topics of the blogs posted by others in users' social hub on users' interests
λ_{F_1}: The impact of the users' relationships on users' interests
λ_{F_2}: The impact of the social hubs' relationships on users' interests.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (31371340) and the National Key Technologies Research and Development Program of China (no. 2016YFB0502604).

References

[1] X. Tang, C. C. Yang, and M. Zhang, "Who will be participating next? Predicting the participation of dark web community," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics 2012, ISI-KDD 2012*, Beijing, china, August 2012.

[2] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization. In NIPS 2008, volume 20".

[3] H. Ma, H. Yang, and M. R. Lyu, "Sorec: social recommendation using probabilistic matrix factorization," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*, pp. 931–940, Napa Valley, Calif, USA, October 2008.

[4] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proceedings of the 4th ACM Recommender Systems Conference (RecSys '10)*, pp. 135–142, Barcelona, Spain, September 2010.

[5] H. Y. Zhang, L. W. Wang, and Y. X. Chen, "Research progress of probabilistic graphical models: a survey," *Journal of Software. Ruanjian Xuebao*, vol. 24, no. 11, pp. 2476–2497, 2013.

[6] G.-F. Sun, L. Wu, Q. Liu, C. Zhu, and E.-H. Chen, "Recommendations based on collaborative filtering by exploiting sequential behaviors," *Ruan Jian Xue Bao/Journal of Software*, vol. 24, no. 11, pp. 2721–2733, 2013.

[7] H. Bao, Q. Li, S. S. Liao, S. Song, and H. Gao, "A new temporal and social PMF-based method to predict users' interests in micro-blogging," *Decision Support Systems*, vol. 55, no. 3, pp. 698–709, 2013.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, 2004.

[9] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the the seventh ACM SIGKDD international conference*, pp. 57–66, August 2001.

[10] R. R. Sinha and K. Swearingen, "Comparing recommendations made by online systems and friends," in *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

[11] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 5–53, 2001.

[12] X. W. Meng, S. D. Liu, Y. J. Zhang, and X. Hu, "Research on social recommender systems," *Journal of Software. Ruanjian Xuebao*, vol. 26, no. 6, pp. 1356–1372, 2015.

[13] L. Guo, J. Ma, Z.-M. Chen, and H.-R. Jiang, "Incorporating item relations for social recommendation," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 37, no. 1, pp. 219–228, 2014.