A new temporal and social PMF-based method to predict users' interests in micro-blogging

Hongyun Bao \(^{a,1}\), Qiudan Li \(^{a,\ast}\), Stephen Shaoyi Liao \(^{b,c,2}\), Shuangyong Song \(^{a,1}\), Heng Gao \(^{a,1}\)

\(^a\) State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
\(^b\) Department of Information Systems and Advanced Transportation Information Systems Research Center, City University of Hong Kong, Hong Kong
\(^c\) Department of Electronic Commerce and Information Management, School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China

Abstract

Micro-blogging is becoming an increasingly popular social media platform where users can discover interesting information about the real world and especially corporations are able to understand customers' demands. The fast diffusion of information and the convenience of micro-blogging have resulted in large audiences sharing their daily activities, exchanging opinions and establishing friendships with others. By analyzing the user-generated contents, one can explore users' potential interests, which helps micro-blogging provide users with better personalized information services. Users' behaviors are affected by opinions of their friends and changes in their interests over time. Based on these intuitions, in this paper we propose a temporal and social probabilistic matrix factorization model to predict users' potential interests in micro-blogging. By exploiting the matrix factorization technique to learn latent features of users and topics, our model analyzes the impacts of time information and users' activities, including posting of tweets and establishing friendships with others, on the latent feature space of users and topics of their interests. The proposed model provides a unified way to fuse the time information and the social network structure to predict users' future interests accurately. The experimental results on Sina-weibo, one of the most popular micro-blogging sites in China, demonstrate the efficiency and effectiveness of our proposed model.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

With the rapid development of the Internet, social media has played an increasingly significant role in our everyday lives, providing reports on world events, improving enterprise influence through social media marketing and so on. Micro-blogging is becoming one of the most popular social media platforms where users can share their daily activities, exchange opinions, publish posts on some trending topics and follow others to get relevant information about their interested topics. If A is following B, B is called A's friend, and A is called B's follower. Thus friendships can either be reciprocated or one-way \(^{[10]}\). The convenience and high frequency of updates of information in micro-blogging have attracted a large number of users to actively participate. For example, Sina-weibo, one of the most popular social media platforms in China, has had over 300 M unique visitors since December 31, 2011 and around 100 M tweets per day.\(^3\) Nowadays, more and more corporations are registering user accounts on Sina-weibo for marketing. For example, Nokia successfully held a product release conference for N8 on August 25, 2010. Corporations utilize micro-blogging not only to introduce their products but also to formulate customer-driven marketing plans by obtaining rich information such as which features customers consider important in their products, new market dynamics and so on. Micro-blogging has become an important e-commerce marketing channel, and the promotion of merchandise is accessible to micro-blogging users around every corner of this platform. Users' interest plays a vital role in the process of micro-blogging's development \(^{[7]}\), which influences the effect of micro-blogging marketing soon afterwards. The research findings in \(^{[5]}\) point out that the accurate prediction of users' interest will improve their satisfaction and promote their buying decisions, which will increase the e-commerce business benefits undoubtedly. Decision makers will also benefit from the interest prediction work; Chen and Cheng and Zhao and Lu \(^{[4,28]}\) proposed that decision makers need to grasp users' interest for raising up their satisfaction and providing reasonable results. Asur and Huberman \(^{[2]}\) tells us that the box-office revenues of movies can be successfully forecasted in advance of their release by analyzing users' interest in micro-blogging. If the forecasted box-office revenues are below expectations, decision makers can provide ways of film promotion with some incentives in time, or some other methods for coming up to their expectations. All in all, it's valuable and meaningful to predict users' interest in social media marketing and so on.

\(^\ast\) Corresponding author. Tel.: +86 10 8261 4835.
E-mail addresses: hongyun.bao@ia.ac.cn (H. Bao), qiudan.li@ia.ac.cn (Q. Li), issiiao@cityu.edu.hk (S.S. Liao), shuangyong.song@ia.ac.cn (S. Song), heng.gao@ia.ac.cn (H. Gao).

1 Tel.: +86 10 62636334.
2 Tel.: +852 34427552.

Available online 24 February 2013
Accepted 4 February 2013
Received in revised form 20 November 2012
Received 15 May 2012
Article history:

Contents lists available at SciVerse ScienceDirect
Decision Support Systems
journal homepage: www.elsevier.com/locate/dss

© 2013 Elsevier B.V. All rights reserved.
media, whether for e-commerce business or decision makers. On one hand, it can help micro-blogging systems provide users with better personalized information and advertising services to motivate users to be more active. On the other hand, corporations can easily capture users’ future interest and make marketing decisions.

In micro-blogging, trending topics are popular topics, which may be related to emerging events and breaking news or topics under the discussion by a large fraction of micro-blogging users [19]. In Sina-weibo, trending topics are edited and complemented, and users are available to enter into the trending topics and take part in the discussion by publishing posts on them. Trending topics often have a clear meaning [11], mainly relating to entertainment, sports, current events and so on. If the user is interested in a trending topic, he/she may publish posts on it. In other words, if a user has published posts on a trending topic, it shows that the user has interest in this topic. Posts published on some trending topics can well reflect users’ interests. Thus, in this paper, we use trending topics to represent users’ interests.

Despite the importance of user interest prediction in micro-blogging, existing works on micro-blogging mainly focus on mining users’ current interests; little work has been done on prediction of users’ potential interests. Nori et al. [20] focused on computing the similarity between a user and a set of resources to predict the user’s interests. However, this method ignores the influence of the user’s friends on his/her interests. Besides, it doesn’t take the evolution of interests into account.

Some researchers have suggested that users are more affected by opinions of their peers than influencers [21,24,25]. By comparing quality of recommendations made by recommender systems to recommendations made by users’ friends, Sinha et al. [23] showed that users’ friends consistently provided better recommendations than recommender systems. In social recommendation, making use of the information in a social graph has recently been receiving increasing research attention. The experimental results [6,9,15–17] show that fusing the social network structure of users with the user–item rating matrix can help make more accurate and personalized recommendations in a social rating network. From this viewpoint, a user’s social network affects users’ behaviors on the Web.

Additionally, interests of Web users change over time. For time-aware recommendation, it is important to capture users’ temporal preferences to make more accurate recommendations. Xiang et al. [12] stated that users’ dynamic preferences are affected by both their long-term and short-term preferences. That means their interests may vary over time. In this case, to capture users’ temporal preferences, it is necessary to follow the evolution of users’ preferences. Generally, recent preferences may play a more important role in predicting current preferences while earlier preferences have relatively smaller contribution to final recommendation. Especially in micro-blogging, the rich information and frequent updates make users’ interests more extensive and changeable over time. Therefore, to improve the accuracy of prediction of users’ interests in micro-blogging, both the social network structure and time information should be taken into consideration.

SocialMF [9] is an effective method for detecting users’ interests by exploiting the matrix factorization techniques and analyzing the influence of users’ friendships on their interests. Based on this model, we propose a temporal and social probabilistic matrix factorization model (TS-PMF) which fuses on social influence and the time information to predict users’ interests in micro-blogging. Following the evolution of users’ interests, to import time information in our model, we make the latent features of users and topics associated with their previous latent features by adopting an exponential time decay function. Using this idea, our approach accurately describes the change of the distribution of the latent feature space of users’ interests. The proposed model can reflect the impacts of users’ interest evolution and users’ friendships on their future interests, thus realizing the prediction of users’ interests. The experimental results on Sina-weibo demonstrate that our model can improve the quality of prediction.

The remainder of this paper is organized as follows. In Section 2, some related work is discussed. Section 3 introduces the proposed model. Results of the detailed experimental analysis are presented in Section 4. Finally, we conclude the paper and present some directions for future work in Section 5.

2. Related work

In this work we propose a novel model to predict users’ interests in micro-blogging. Our work is related to prediction of user interest in micro-blogging, trust-aware recommendation and time-aware recommendation. In this section we review the related works.

2.1. User interest analysis and prediction

Banerjee et al. [3] gathered tweets data from Twitter across ten (worldwide) cities over a period of four weeks to generate an exhaustive list of keywords and then applied statistical and mining techniques to discover the distribution of users’ interest on categories such as “movie”, “food”, “game”, “dinner” and so on. Xu et al. [30] proposed a modified author-topic model to discover users’ topics of interest on Twitter by filtering out interest-unrelated tweets (noisy posts) from the aggregated user profiles. These studies concentrated on the text level analysis of user interests. Yan et al. [27] established a human dynamic model co-driven by interest and social identity and showed that users’ interest in sending posts is positively correlated with the number of comments on their previous posts.

Existing works on micro-blogging mainly focus on mining users’ current interests; little work has been done on prediction of users’ potential interests. Nori et al. [20] proposed ActionGraph, a new graphic representation for modeling users’ multinomial, time-evolving actions, to compute the similarity between a user and a set of resources to predict the user’s interest. ActionGraph is a bipartite graph whose edge connects an action node at some point in time and the object nodes representing users and resources. It preserves the time information for each user by representing the same action in different times as different action nodes. However, it ignores the influence of the user’s friendships on his/her interest. Besides, it doesn’t take the evolution of interest into account.

2.2. Trust-aware recommendation

In social networks, users can follow others whom they are interested in, and then they may have social interactions or connections instead of being independent and identically distributed. Many researchers have recently focused on trust-aware recommender systems. Ziegler and Golbeck [29] established two frameworks for investigating and analyzing the correlation between interpersonal trust and interest similarity, and empirical results showed that the mean similarity of trusting and trusted peers exceeded the arbitrary user similarity. Massa and Avesani [18] show that the idea of Trust-aware Recommender System is not to search for similar users as CF (Collaborative Filtering) does but to search for trustable users by exploiting trust propagation over the trust network. The items appreciated by these users are then recommended to the active user. They present a complete evaluation of Trust-aware Recommender System, by comparing different algorithms, ranging from traditional CF ones to algorithms that utilize only trust information with different trust metrics and algorithms that combine both trust and similarity to baseline algorithms. Those methods are all memory-based methods not scalable to very large datasets.

Ma et al. [6,15,16] studied the relationship between the trust network and the user–item matrix systematically and proposed the methods integrating social network structure and the user–item rating matrix, which were based on probabilistic factor analysis. Jamali and Ester [9] stated that the real world recommendation processes are not reflected in the model [6,15,16]. Due to social influence,
related people in a social network influence each other to become more similar. They proposed a SocialMF model by incorporating trust propagation into a matrix factorization for recommendation in social network. Their experimental results demonstrate that SocialMF outperforms existing methods for social network based recommendation.

In the advertising recommendation system of micro-blogging, some researchers are taking advantage of friendships of users [14].

2.3. Time-aware recommendation

Ding and Li [26] presented a new time weight collaborative filtering algorithm using an exponential time decay function to compute time weights for different items according to each user and each cluster of items. Xiang et al. [12] argued that user preferences often exhibit long-term and short-term factors and proposed a STG model to capture users' dynamic preferences by considering all items viewed by a user as his long-term preferences and items viewed by him at a given time as his short-term preferences. Xiong et al. [13] presented a Bayesian Probabilistic Tensor Factorization algorithm for modeling evolving relational data by organizing the ratings into a three-dimensional tensor whose three dimensions correspond to user, item and time slices, assuming each time feature vector depends only on its immediate predecessor. Ahmed et al. [1] argued user profiles were temporal and changed user activity patterns, thus presenting a comprehensive statistical framework for user profiling based on topic models. Their method modeled topical interests of a user dynamically where both the user association with the topics and the topics themselves were allowed to vary over time, ensuring that the profiles remain current.

Thus, in extant literature, little work has been done on prediction of users' potential interests in micro-blogging; works on trust- and time-aware recommendation show that both the social network structure and the time information are important for recommendation in social network. In this paper, we propose a TS-PMF model to predict users' interests in micro-blogging, which provides a unified way to integrate the social network structure and the time information. Specifically, we express users' interests as a series of temporal matrices and use the probabilistic matrix factorization technique to learn the users' latent feature space and topics' latent feature space by employing users' social network and temporal matrices.

3. The proposed user interest prediction model

In this section, first we introduce the theoretical background for our proposed model. Second, we illustrate how to fuse social network structure and time information in our TS-PMF model to predict users' interests in micro-blogging.

3.1. Theoretical background

We introduce some notations first. We have a set of users $U = \{u_1, ..., u_u\}$ and a set of topics $Z = \{z_1, ..., z_z\}$ in a micro-blogging dataset. We construct a user–topic matrix $R \in \mathbb{R}^{u \times z}$ to represent users' interests, where we set $R_{ij} = 1$ if $u_i$ has published posts on $z_j$. In micro-blogging, each user can follow others whom he is interested in. Then users' friendships can be described as a user–user matrix $C \in \mathbb{R}^{u \times u}$, where $C_{ij} = 1$ that denotes $u_i$ has followed $u_j$. Furthermore, we record the set of $u_i$'s friends as $N(i)$. The task of predicting users' interests is to predict the relational scores for a given user $u_i$ on topics $Z = \{z_1, ..., z_z\}$ in the future using $R$ and $C$. Salakhutdinov and Mnih [22] have shown that it is very effective to employ matrix factorization techniques to learn the latent characteristics of users and topics and predict the scores using these latent characteristics. Let $U \in \mathbb{R}^{d \times u}$ and $V \in \mathbb{R}^{d \times z}$ be the latent user and topic feature matrices, with column vectors $U_i$ and $V_j$ representing $d$-dimensional user- and topic-latent feature vectors of $u_i$ and $z_j$, respectively. The goal of matrix factorization is to model each score as the production of user- and topic-latent feature vectors, i.e. $R_{ij} \approx U_i^T V_j$, where $U_i^T$ is the transpose of $U_i$. As is shown in the SocialMF model, the conditional probability of the known scores is defined as:

$$p(R|U,V,\alpha_R^2) = \prod_{i=1}^m \prod_{j=1}^n \prod_{k=1}^{N(j)} N(R_{ij}|g(U_i^T V_j, \alpha_R^2))^{\delta_{ij}}$$

(1)

where $N(x|\mu, \sigma^2)$ is the Gaussian distribution with mean $\mu$ and variance $\sigma^2$, and $\delta_{ij}$ is the indicator function that is equal to 1 if $R_{ij} = 1$ and equal to 0 otherwise. The function $g(x)$ is the logistic function $g(x) = 1 / (1 + \exp(-x))$, which makes it possible to bound $x$ within the range [0,1].

Social network researchers have pointed out that the social network structure plays an important role in users' behavior [21,23,25]. Specifically, a user is more and more similar to his/her friends. From this perspective, SocialMF incorporates social influence into the matrix factorization for recommendation in social network [9].

Because the behavior of a user $u_i$ is affected by his/her friends $N(i)$, the latent feature vector of $u_i$ is dependent on latent feature vectors of all his/her friends $u_v \in N(i)$. For user latent features, there are two factors: the zero-mean Gaussian prior and the latent features of his/her friends. Therefore,

$$p(U|\alpha_U^2, \alpha_Z^2) \propto p(U|\alpha_U^2) \times p(\alpha_U^2)$$

(2)

$$= \prod_{i=1}^m N(U_i|0, \alpha_U^2) \times \prod_{i=1}^m \left[ \prod_{v \in N(i)} C_{iv} U_v, \alpha_Z^2 \right]$$

where each row of $C$ is normalized, through $C_{iv} = 1/|N(i)|$ with $v \in N(i)$.
Now, the posterior probability of latent variables $U$ and $V$ can be obtained through a Bayesian inference. Maximizing the log of the posterior distribution is equivalent to minimizing the following sum-of-squared-errors objective function with quadratic regularization terms:

$$E(R, U, V) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (R_{ij} - g(U_i^T V_j))^2 + \frac{\lambda_U}{2} \sum_{i=1}^{m} U_i^T U_i + \frac{\lambda_V}{2} \sum_{j=1}^{n} V_j^T V_j + \frac{\lambda_C}{2} \sum_{i=1}^{m} (U_i - \sum_{v \in N(i)} C_{iv} U_v)^T (U_i - \sum_{v \in N(i)} C_{iv} U_v).$$

The above optimization can be done efficiently using gradient descent. The graphical model for SocialMF is presented in Fig. 1.

3.2. Proposed user interest prediction model

In the near future, micro-blogging users may focus on new topics, and they may also show different concerns to the topics which they have been interested in for a period of time. The above users' dynamic interests are mainly affected by their friends and their historical favorites. Based on these intuitions, the primary motivation of our model is to provide a unified way to fuse the social network structure and the evolution of users' interests for predicting users' interests accurately. In this section, we extend SocialMF to predict users' future interests in micro-blogging by taking time information into account.

3.2.1. Toy example

We use a simple toy example to demonstrate the proposed model for predicting users' interests in micro-blogging. There are 6 users $U = \{u_1, ..., u_6\}$ and 8 topics $Z = \{z_1, ..., z_8\}$ in total. The relationships among users (nodes) are illustrated in Fig. 2(a), and the edge from $u_i$ to $u_j$ denotes that $u_i$ has followed $u_j$. We have segmented users' historical data into 3 time points ($T_1, T_2, T_3$). As shown in Fig. 2(b, c, d), each user has published posts on some topics in $T_t$ ($t = 1, 2, 3$) to express interest in those topics. Our main goal is to predict users' interests in the future time $T_4$. As elaborated in Section 1, a user's social friendships make his/her interest similar to his/her friends and the current interest is also affected by historical interest. Therefore, we minimize the sum-squared distance to the target matrix $R_t$ by $U_t^T V_t$ to factorize...
affected by his friends who had focused on this topic for a while. The users’ dynamically changing interests can be expressed as the collection factors may prompt him to perform like this; maybe he was informed of the message that the new iPad has been released, or he was likely interested in T. And then according to the values in R, our model predicts evolution of interest, our model predicts feature space and the current user was fond of some electronic products in early days, but lately, he took a great interest in the topic of new iPad has focused on this, and only u1 has paid attention to z6 in T1, our model provides z6 for u4 in the second place in the prediction list. While both u1 and u2 are interested in z2 in T5, and u4 has never published posts on z2, our model places z2 in the third place.

3.2.2. The proposed model

It’s noticeable that users’ interests are changing over time; users show their concerns to different topics at different times. For example, a user was fond of some electronic products in early days, but lately, he took a great interest in the topic of new iPad for a few days. Two influencing factors may prompt him to perform like this; maybe he was informed of the message that the new iPad has been released, or he was likely affected by his friends who had focused on this topic for a while. The users’ dynamically changing interests can be expressed as the collection

the current user–topic matrix R, fusing the social network structure and the evolution of users’ interest, where Ui denotes the user latent feature space and Vt represents the topic feature space in time t. If we use 5 dimensions to perform the matrix factorization, we obtain Ut and Vt (t = 1, 2, 3) and then compute the mean matrices Mu and MV of Ut and Vt for predicting the user-matrix R4 in T2:

\[
M_{u_4} = \begin{bmatrix}
-0.3241 & -0.2288 & -0.8033 & -0.3353 & -0.4104 & -0.6743 \\
-0.8120 & -0.2772 & -1.1198 & -0.4928 & -0.8098 & -0.6044 \\
-0.8709 & 0.2190 & -0.0806 & -0.0979 & -0.4863 & 0.2643 \\
-0.3309 & -1.1181 & 0.8569 & -0.6012 & -0.1000 & 0.3677 \\
-0.5246 & -0.6829 & 0.3100 & 0.2823 & -0.9464 & 0.5734
\end{bmatrix}
\]

\[
M_{v_4} = \begin{bmatrix}
-0.5594 & -0.4758 & -0.5559 & 0.2031 & -0.2315 & -1.0572 & -0.2741 & -0.2590 \\
-0.6038 & -0.6740 & -0.6816 & -0.0477 & -0.3540 & -0.8846 & -0.8765 & -0.7627 \\
0.0129 & -0.1069 & -0.0157 & -0.4844 & -1.988 & -0.0478 & -0.5087 & -0.0691 \\
0.7660 & -0.6818 & 0.8971 & -0.5544 & -0.6918 & 0.5148 & 0.1265 & -0.8164 \\
0.2424 & -0.9689 & 0.3805 & -0.4788 & -0.0656 & 0.6503 & -0.0563 & -0.8368
\end{bmatrix}
\]

where M_{u_{4,j}} and M_{v_{4,j}} are the column vectors and denote the latent feature vector of u4 and topic zj, respectively, in T4. Then we use R4≈M_{u_{4,j}}M_{v_{4,j}} to predict all the values in R4, where we need to transfer the value of M_{u_{4,i}}M_{v_{4,j}} using the function g(x) introduced in Section 3.1. And then according to the values in R4 as shown in Fig. 2(e), each user is provided a topic list he/she is likely to prefer in the future.

We take u4 as an example to explain the reasoning behind the prediction by our model. As shown in Fig. 2(a), u4 has followed u1 and u2, u4 has focused on z3 in T2 and u3 has also focused on z3 in T2 and T3 as well as u1 does in T3. Considering the impacts of his/her friends and the evolution of interest, our model predicts u4 will be interested in z3 in T4 with the maximum probability of 0.67. Since u4 is interested in z6 in T3, and only u1 has paid attention to z6 in T1, our model provides z6 for u4 in the second place in the prediction list. While both u1 and u2 are interested in z2 in T5, and u4 has never published posts on z2, our model places z2 in the third place.

Fig. 3. The framework of predicting users’ interest.
of users' sequential interest matrix at different times, each of which is constructed as a temporal user–topic matrix $R_t \in \mathbb{R}^{m \times n}$, where $t$ ($t = 1, 2, ..., N$) is the time label of the data section. In the user–topic matrix, the value of element $R_{ij}$ equals 1 means that the user $u_i$ is interested in the topic $j$ at time point $t$. Meanwhile, users' friendships can be expressed as a user–user matrix $C$. Our proposed model is designed to utilize users' sequential interest matrices $[R_1, ..., R_N]$ at the existing time points ($t = 1, 2, ..., N$) and the users' friendships matrix $C$, to predict users' interest in the near future.

The framework of our proposed model is shown in Fig. 3. At each time point, our model will find the matrix $R_t \approx U_t^i V_t$ ($t = 1, ..., N$) which minimizes the sum-squared distance to the target matrix $R_t$ under the constraints of the temporal and social impacts, where $U_t^i$ and $V_t$ represent the users' and topics' latent feature spaces in time $t$.

In time $t$, the conditional distribution probability of the observed items in $R_t$ is similar to that in Eq. (1):

$$p(R_t|U_t, V_t, \sigma^2_{R_t}) = \prod_{i=1}^{m} \prod_{j=1}^{n} \left[ N(R_{ij}|g(U_t^i, V_t), \sigma^2_{R_t}) \right]^{R_{ij}}.$$

(4)

$U_t^i$ represents the latent feature vector of user $u_i$ in time $t$. The user's latent feature vector follows Gaussian distribution [6,9,15,16], which is decided by the mean value of Gaussian distribution. Changes in mean values always reflect changes in users' interests. Normally, users' current interests will be affected by his historical favorites; the influence will be greater with the closer interest to the current time. Amazingly, the exponential decay function can describe the influence process effectively [26], with the following mathematical expression:

$$f(k) = \exp\left(-\frac{t-k}{\beta}\right) (k \in \{1, 2, 3, ..., t-1\} \beta > 0).$$

(5)

In Eq. (5), the value of $\beta$ presents the kernel parameter, and the value of $t - k$ shows us the time interval between the $k$-th time point and the current time $t$. The higher value $t - k$ presents the earlier time $k$ from current time $t$, accompanied with the smaller influencing value $f(k)$. It's evident that the exponential decay function can vividly influence the user's historical favorites have on the current interests.

Based on the above analyses, we utilize exponential decay function with kernel parameter $\beta$ to compute the mean value matrix of user-latent feature and the mean value matrix of topic-latent feature in time $t$. The computing formulation is listed below:

$$M_{U_t} = \theta \sum_{k=1}^{t-1} \exp\left(-\frac{t-k}{\beta}\right) U_k, M_{V_t} = \theta \sum_{k=1}^{t-1} \exp\left(-\frac{t-k}{\beta}\right) V_k$$

(6)

where $M_{U_t}, M_{V_t}$ are the mean matrices of $U_t$ and $V_t$ with spherical Gaussian priors, and $\theta$ is a weight parameter that indicates how important the whole previous time points are to the current one.

In summary, the user's latent feature vector is affected by two factors, of which are the latent feature vectors of his historical interests and his friends' interests. Therefore, the conditional distribution probability of users' latent features can be expressed like this:

$$p(U_t|\{R_1, R_2, ..., R_{t-1}\}, C, \sigma^2_{U_t}, \sigma^2_{R_t}) \propto$$

$$p(U_t|\{R_1, R_2, ..., R_{t-1}\}, \sigma^2_{U_t}) \times p(U_t|C, \sigma^2_{R_t})$$

$$= \prod_{i=1}^{m} N(U_{ij}|M_{U_{ij}}, \sigma^2_{U_t}) \times \prod_{i=1}^{m} N(U_{ij}|\sum_{v \in N(i)} C_v U_{iv}, \sigma^2_{R_t}).$$

(7)

The above normal distribution consists of two parts. The first part, that is $\prod_{i=1}^{m} N(U_{ij}|M_{U_{ij}}, \sigma^2_{U_t})$, informs us of the fact that the distribution vectors of users' latent feature space are close to their historical distribution vectors. The second part, which is $\prod_{i=1}^{m} N(U_{ij}|\sum_{v \in N(i)} C_v U_{iv}, \sigma^2_{R_t})$, tells us that the distribution vectors of users' latent feature space are also close to their friends' distribution vectors.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Summary of notations used in this paper.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_t$</td>
<td>The user–topic matrix in time $t$</td>
</tr>
<tr>
<td>$C$</td>
<td>The user–user matrix</td>
</tr>
<tr>
<td>$U_t^i$</td>
<td>The user’s latent feature space in time $t$</td>
</tr>
<tr>
<td>$V_t$</td>
<td>The topics’ latent feature space in time $t$</td>
</tr>
<tr>
<td>$M_{U_t}$</td>
<td>The mean matrix of $U_t$ with spherical Gaussian priors in time $t$</td>
</tr>
<tr>
<td>$M_{V_t}$</td>
<td>The mean matrix of $V_t$ with spherical Gaussian priors in time $t$</td>
</tr>
<tr>
<td>$\theta$</td>
<td>A weight that indicates how important the whole previous time points are to the current one</td>
</tr>
<tr>
<td>$\beta$</td>
<td>The kernel parameter</td>
</tr>
<tr>
<td>$d$</td>
<td>The dimension of latent feature space</td>
</tr>
<tr>
<td>$\lambda_{C}$</td>
<td>The impact of the social network on users' interest</td>
</tr>
<tr>
<td>$\lambda_{U}$</td>
<td>The impact of the users' latent feature vectors on users' interest</td>
</tr>
<tr>
<td>$\lambda_{V}$</td>
<td>The impact of the topics' latent feature vectors on users' interest</td>
</tr>
<tr>
<td>$T$</td>
<td>The days for partitioning the data set</td>
</tr>
</tbody>
</table>
After that, through a Bayesian inference, we have the following equation for the posterior probability over latent features of users and topics,

\[ p(U_t, V_t | R_1, R_2, \ldots, R_{t-1}, R_t, C, \sigma_u^2, \sigma_v^2, \alpha_{U_t}, \alpha_{V_t}) \propto \]

\[ p(R_t | U_t, V_t, \sigma_v^2) p(U_t | C, \sigma_u^2) p(U_t | R_1, R_2, \ldots, R_{t-1}, \alpha_{U_t}) p(V_t | R_1, R_2, \ldots, R_{t-1}, \alpha_{V_t}). \]  

(8)

The log of the posterior distribution for our proposed model at time point \( t \) is given by

\[
\ln p(U_t, V_t | R_1, R_2, \ldots, R_t, C, \sigma_u^2, \sigma_v^2, \alpha_{U_t}, \alpha_{V_t}) =
\]

\[- \frac{1}{2\sigma_u^2} \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} g(U_t^j, V_t^j)^2 - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} \ln \sigma_u^2 \]

\[- \frac{1}{2\sigma_v^2} \sum_{i=1}^{m} (U_t - M_{0,t})^T (U_t - M_{0,t}) - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} (V_t - M_{0,t})^T (V_t - M_{0,t}) \]

\[- \frac{1}{2} \sum_{i=1}^{m} \left( U_t - \sum_{v=1}^{N(t)} C_{w} U_{t,v} \right)^T \left( U_t - \sum_{v=1}^{N(t)} C_{w} U_{t,v} \right) \]

\[- \frac{1}{2} \left( \ln \sigma_{V_t}^2 + \ln \sigma_{U_t}^2 + \ln \sigma_{V_t}}{2} + W. \right)

(9)

In this equation, \( W \) is a constant. Once the parameters \( \sigma_u^2, \sigma_v^2, \alpha_{U_t}, \alpha_{V_t} \) are fixed, 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:

\[
E(U_t, V_t; (R_1, R_2, \ldots, R_t), C) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} r_{ij} g(U_t^j, V_t^j)^2 + \frac{\lambda_u}{2} \left\| U_t - M_{0,t} \right\|_F^2 \]

\[+ \frac{\lambda_v}{2} \left\| V_t - M_{0,t} \right\|_F^2 + \frac{\lambda_c}{2} \sum_{i=1}^{m} \left( U_t - \sum_{v=1}^{N(t)} C_{w} U_{t,v} \right)^T \left( U_t - \sum_{v=1}^{N(t)} C_{w} U_{t,v} \right). \]

(10)

In Eq. (10), \( \lambda_u = \partial E / \partial \sigma_u^2 \), \( \lambda_v = \partial E / \partial \sigma_v^2 \), \( \lambda_c = \partial E / \partial \alpha_{U_t} \), \( \lambda_{V_t} = \partial E / \partial \alpha_{V_t} \), and \( \| \cdot \|_F \) are the Frobenius norm. The first sum in Eq. (10) represents the sum-squared error between \( R_i \) and \( U_t^jV_t \), while the second term and the third term denote the sum-squared distance from the users’ and the topics’ latent feature space to the prior ones, we utilize the forth sum to represent the sum-squared deviation between the user’s latent feature space to the prior one.
space and his friends’ latent feature space. We can find a local optimal value of the objective function in Eq. (10) by performing gradient descent in \(U_t\) and \(V_t\); that is,

\[
\frac{\partial E}{\partial U_{ij}} = \sum_{t=1}^{n} \rho g' \left( \left(U_t, V_{ij}^t\right) \left(g(U_t, V_{ij}^t) - R_t, ij\right) V_{ij} + \lambda_0 \left(U_t - M_{U_{ij}}\right) \right) + \lambda_0 \left(U_t - \sum_{c=0}^{C_{16}} C_{uv} U_{ct}^tV_{ij}^t\right) \tag{11}
\]

and

\[
\frac{\partial E}{\partial V_{ij}} = \sum_{t=1}^{m} \rho g' \left( \left(U_t, V_{ij}^t\right) \left(g(U_t, V_{ij}^t) - R_t, ij\right) U_{ij} + \lambda_0 \left(V_t - M_{V_{ij}}\right) \right) \tag{12}
\]

where \(g'(x) = \exp(-x) / (1 + \exp(-x))^2\) is the derivative of logistic function \(g(x)\). In order to reduce the model complexity, in all experiments we set \(\lambda_0, \lambda_1 = \lambda_U = \lambda_V = \lambda\).

Our proposed model provides an effective method to predict users’ interests by integrating users’ friendships and users’ historical interests. Our model can effectively take various factors that will have influence on users’ interests change into account, thus achieving accurate forecasts of users’ future interests. The process of predicting users’ interests with our model is described in Algorithm 1. And the notations used throughout the paper are summarized in Table 1.

### Algorithm 1. The process of predicting users’ interests

**Input:**
- dataset: \(\{R_t, \ldots, R_N\}\), \(C\)
- the dimension of the latent feature: \(d\)
- parameters: \(\lambda_r, \lambda_c, \lambda_\theta, \theta, \beta\)
- an updating parameter: \(\delta\)
- convergence parameter: \(\varepsilon\)
- the maximum number of iterations: \(K\)

**Output:**
- the user–topic matrix in time segment \(N+1:R_{N+1}\)

**Learning:**
- set \(M_{u_0} = \text{zeros}(d, n)\), \(M_{v_0} = \text{zeros}(d, m)\).
- For \(t = 1, \ldots, N\)
  - Step 1: initialize \(U_t, V_n\), \(U_{t0} = U_n, V_{t0} = V_n, E_0 = \inf\).
  - If \(t > 1\)
    - compute the mean matrices \(M_u\) and \(M_v\) in Eq. (6).
  - End If
  - For \(l = 1, \ldots, K\)
    - Step 2: compute the gradient descent in Eq. (11) and (12).
    - Step 3: updating:
      \[
      U_t = U_t + \delta \frac{\partial E}{\partial U_t}, V_t = V_t + \delta \frac{\partial E}{\partial V_t} \tag{13}
      \]
    - Step 4: compute \(E_t\) in Eq. (10)
      - If \(E_t < \varepsilon\), then break out of the loop.
      - Else get \(E_0 = \min\{E_0, E_t\}\), if \(E_0 = E_t\), then set \(U_{t0} = U_t, U_{t0} = V_t\).
  - End If
  - Step 5: \(U_{t} = U_{t0}, V_{t} = V_{t0}\).
  - End For
  - Step 6: compute \(M_{u}\) and \(M_{v}\) in Eq. (6).
  - Step 7: predict \(R_{t,ij}\) using \(R_{t,ij} = M_{u}^t M_{v}\).

### 3.3. Complexity analysis

The main computation of learning parameters involves evaluating the object function \(E\) in Eq. (10) and its gradients against variables \(U_t\) and \(V_t\) in Eqs. (11) and (12), including computing the mean matrices \(M_{u}\) and \(M_{v}\) in Eq. (6). In each time segment \(t\), the complexity of evaluation \(E\) is \(o(\rho_R + \rho_C d)\), where \(\rho_R\) and \(\rho_C\) are the numbers of nonzero entries in matrices \(R_t\) and \(C\), respectively. The cost of computing the

| Performance on different users of our proposed model TS-PMF (\(d = 45\)) |
|-----------------|---|---|---|---|---|---|---|---|---|
|                | 1–10 | 11–20 | >20 |
| PMF            | Pr1 | 0.2684 | 0.2711 | 0.3919 | 0.3730 | 0.4878 | 0.4878 | 0.7439 | 0.6316 | 0.6316 | 0.8684 | 0.9211 |
| SocialMF       | Pr3 | 0.2026 | 0.2079 | 0.2261 | 0.2207 | 0.5000 | 0.5163 | 0.5813 | 0.6870 | 0.7105 | 0.7281 | 0.8158 | 0.8947 |
| T-PMF          | Pr5 | 0.1642 | 0.1621 | 0.1627 | 0.1627 | 0.4854 | 0.4627 | 0.4780 | 0.5229 | 0.7105 | 0.7073 | 0.7632 | 0.7895 |
| TS-PMF         | Pr10| 0.1039 | 0.1068 | 0.1027 | 0.1051 | 0.3659 | 0.3683 | 0.3439 | 0.3598 | 0.6184 | 0.6079 | 0.6132 | 0.6211 |
gradients is \( o(\rho_c d + \rho_d d/m) \). Then all cost in one iteration is \( o(\rho_k d + \rho_c d + \rho_d d/m + (N-1)/2) \). Thus, our model is effective for handling large datasets when \( R_t \) and \( C \) are sparse.

4. Experimental analysis

In this section, comprehensive and systematic analyses are conducted to evaluate the proposed user interest prediction model. The process of collection of the dataset used in the empirical work and the evaluation metrics are presented first. Next, we explain the purpose of our experiments in detail. Finally, the performance of our model is compared with results of the other four models; results of the comparison verify the efficacy of the proposed model.

4.1. Description of the Sina-weibo dataset

In this paper, we use Sina-weibo API to gather users’ following links and data of topics of their interest from Oct 29, 2011 to Nov 13, 2011. In this data set, a timestamp is available for each user. After removing users with less than 16 posts, we had 1170 users and 2788 topics. In this dataset, each user has on average 1.57 expressed topics per day and data of topics of their interest from Oct 29, 2011 to Nov 13, 2011. In this data set, a timestamp is available for each user. After removing users with less than 16 posts, we had 1170 users and 2788 topics. In this dataset, each user has on average 1.57 expressed topics per day and 2.04 following links. We can observe that both the user–user matrix and the user–topic matrix are very sparse.

In this paper, the first 15 days’ data is used for training and the last day’s data for testing. Because the average time of duration of each topic is 2.72 days in our data set, we set \( T = 3 \) days, and accordingly get \( N = 5 \).

4.2. Evaluation metric

Prediction results are evaluated by ranking the topics for all users according to the scores in \( R_{W+1} \). Since users are concerned about the top ranking topics, the metrics of precision in top-\( n \) is, therefore, adopted to measure the prediction quality of our proposed approach in comparison with other methods, defined as:

\[
Pr_{n} = \frac{N_{corr}(n)}{N_u \times n}
\]  

(14)

where \( Pr_{n} \) is the precision in top-\( n \), and \( N_u \) is the number of users in the testing data set. We consider the correct topics of a user as those which appear in users’ future posts in the testing data set. \( N_{corr}(n) \) is the number of correct topics in top-\( n \) prediction list for users who are in the testing data set.

4.3. Purpose of our experiments

Our model is based on the intuition that both the social network structure and the evolution of users’ interest affect users’ future interest. Our experiments are intended to address the following questions:

1. Do the social network structure and the evolution of users’ interest have impact on users’ interest in micro-blogging in the future?
2. How do model parameters \( \lambda_c \) and \( \theta \) affect the accuracy of prediction?
3. How does our model select an appropriate dimension of latent feature space \( d \)?
4. Does the division of temporal sections \( T \) affect the resulting performance?
5. Is our model effective for active users, who help improve the social influence of micro-blogging?

To answer these questions, we proceed in the following fashion. For the first question, we conduct a set of experiments to separately take the social network structure and the evolution of users’ interest into account for proving the effectiveness of these factors in shaping future interest of users in micro-blogging. For the second question, we illustrate the impacts of parameters with different values. For the third question, we will perform some further experiments to select the appropriate dimensionality of our proposed model by minimizing the accumulated error. For the fourth question, we will take further experiments to analyze the impact of division of temporal sections on the resulting performance.

The social influence of micro-blogging is dependent upon fast fusion of information and the rich user-generated content. Therefore, active users who are keen on publishing posts on trending topics are important for micro-blogging. For the fifth question, it intuitively shows that the performance of active users plays an important role in prediction.

4.4. Experimental results

In this sub-section, to demonstrate the usefulness and effectiveness of our proposed model TS-PMF, we compare it with four other models:

1. Probabilistic matrix factorization (PMF): This is the baseline matrix factorization approach proposed in [22], which only uses the user–topic matrix without temporal information.
2 SocialMF: This is the model proposed in [9], which takes the social network into account and uses the user–user matrix and user–topic matrix without temporal information. We set $\lambda_c = 0.001$ for SocialMF in our experiments, which is the optimum value according to the best performance on our data.

3 Temporal probabilistic matrix factorization (T-PMF): This is the model using a series of temporal matrices, in which we just import temporal impact on users’ interest into the PMF model.

4 MT1: This is the model presented in [18], using only the users explicitly followed by the target user.

In all the experiments, some parameters setting of our approach are $\beta = 3$, $\lambda_0 = \lambda_\nu = 0.0001$. And we take the users who have published posts at least on one topic on the 16th day for testing. Other parameters are set as $\lambda_c = 0.001$ and $\theta = 0.2$, which are explained in Sections 4.6 and 4.7.

We perform some further experiments to select the appropriate dimensionality of our proposed model by minimizing the accumulated error value, as shown in Fig. 4. In those experiments, we range the dimension of latent feature space from 5 to 100 by an interval of 5, and find the best performance at $d = 45$, eventually. And we show the experimental results with the dimensionality $d = 10, d = 45$ and $d = 50$ in Table 2, where we indicate the best performance of those models in bold type, with the same dimensionality $d$. From Table 2, we can observe that with different dimensionalities, both SocialMF and T-PMF improve the accuracy of prediction in comparison with PMF. That is, friendships and temporal factor are both useful for user interest prediction in micro-blogging. And T-PMF can also improve the PMF. That is, friendships and temporal factor are both useful for user interest.

TS-PMF makes the accuracy increase in comparison with the results of those models in bold type, with the same dimensionality $d$. From Table 2, we can observe that with different dimensionalities, both SocialMF and T-PMF improve the accuracy of prediction in comparison with PMF. That is, friendships and temporal factor are both useful for user interest prediction in micro-blogging. And T-PMF can also improve the PMF. That is, friendships and temporal factor are both useful for user interest.

4.5. Performance on different users

Users are grouped into 3 classes: “1–10”, “11–20” and “>20”, denoting on how many topics they have published posts on the 16th day. The experimental results for different users with dimensionality $d = 45$ are shown in Table 3, where we indicate the highest accuracy of those models in bold type, with the same user class; T-PMF outperforms other models on users with the label “1–10”, and TS-PMF makes the accuracy increase in comparison with the results of $d = 45$, although they still outperform PMF and SocialMF. That means only increasing $d$ in a certain range can improve the accuracy of prediction results by our proposed model.

4.6. Impact of $\lambda_c$ on the results

Parameter $\lambda_c$ in Eq. (6) is a weight that indicates how important the previous time points are to the current one, for user-latent feature matrix and topic-latent feature matrix. If $\theta = 0$, our model is the same as T-PMF, which takes only time information into account, and if $\theta = 1$, we consider that the evolution of users’ interest plays a decisive role in the current users’ latent feature space and topics latent feature space. Fig. 6 shows the influence of $\theta$ for users publishing posts at least on one topic on the 16th day when $d = 45$. We observe that values of $\theta$ affect the accuracy of predicting users’ interests. From Fig. 6, we can see TS-PMF has its best result for $\theta = 0.2$.

4.7. Impact of $\theta$ on the results

Parameter $\theta$ in Eq. (6) is a weight that indicates how important the previous time points are to the current one, for user-latent feature matrix and topic-latent feature matrix. If $\theta = 0$, our model is the same as T-PMF, which takes only time information into account, and if $\theta = 1$, we consider that the evolution of users’ interest plays a decisive role in the current users’ latent feature space and topics latent feature space. Fig. 6 shows the influence of $\theta$ for users publishing posts at least on one topic on the 16th day when $d = 45$. We observe that values of $\theta$ affect the accuracy of predicting users’ interests. From Fig. 6, we can see TS-PMF has its best result for $\theta = 0.2$.

4.8. Impact of $T$ on the results

Parameter $T$ denotes the days for partitioning the data set. In this paper, we set $T = 3$ because of the average time of duration of each topic in our dataset. In addition, we take further experiments to analyze the impact of $T$ on the resulting performance. As is shown in Table 2, we can see that values of $\lambda_c$ for users publishing posts at least on one topic on the 16th day when $d = 45$. As shown in Fig. 5, TS-PMF obtains the best performance for $\lambda_c = 0.001$.

4.9. Impact of $d$ on the results

Different values of $d$ should add more flexibility to the models and improve the results. However, comparing results in Table 2, with the dimensionality increasing more than 45, neither T-PMF nor TS-PMF makes the accuracy increase in comparison with the results of $d = 45$, although they still outperform PMF and SocialMF. That means only increasing $d$ in a certain range can improve the accuracy of prediction results by our proposed model.

4.10. Impact of different values of $\lambda_c$ on the results

Parameter $\lambda_c$ in Eq. (6) is a weight that indicates how important the previous time points are to the current one, for user-latent feature matrix and topic-latent feature matrix. If $\theta = 0$, our model is the same as T-PMF, which takes only time information into account, and if $\theta = 1$, we consider that the evolution of users’ interest plays a decisive role in the current users’ latent feature space and topics latent feature space. Fig. 6 shows the influence of $\theta$ for users publishing posts at least on one topic on the 16th day when $d = 45$. We observe that values of $\theta$ affect the accuracy of predicting users’ interests. From Fig. 6, we can see TS-PMF has its best result for $\theta = 0.2$.
in Fig. 7, the X-axis represents the days for partitioning our data set, including 2, 3, 4 and 5 days respectively, and the Y-axis represents the precision of predicting users’ interest in the test data. And we can observe that when we set \( T = 3 \) days for partitioning data sets, our proposed model achieve the best performance when \( d = 45 \).

4.9. Prototype system

Fig. 8 shows our prototype system interfaces, which allow users to see not only a list of topics on which a given user has published posts in the past and the ones he will prefer in the future, but also the related posts on the given topic, which appears in the user’s topic list.

The upper graph of Fig. 8 gives an example of searching a given user “The vagrant 1885838067”. His interests include costume drama, comic, movie, daily life, Korean current star and so on. Since he focused on costume drama “Introduction of the Princess” and “Each step Escape” for a long time, our model predicts he may continue to prefer them as well as other models. Meanwhile, some of his interests have changed over time. For example, he only paid attention to Korean Current star such as “Girls’ Generation” and “U-know” at first, and then he showed his interests in TV series of stars such as “Poseidon”. Therefore, our model provides him a TV series “Skip beat” of Korean Current star. Furthermore, he has published posts on “Nokia N9”, “BYD G3”, “Gas station”, “Air quality”, “Highway”, a sightseeing spot named “Zhang jiajie” and “Good trip”. This shows he is interested in some topics about travelling by car. Additionally, some of his friends have been paying attention to a topic named “Posting tweets with mobile on trips”. Our model offers “Posting tweets with mobile on trips” to him. From the above example, we can see that our proposed method can effectively detect users’ future interests by considering the social network structure and the evolution of users’ interest.

The bottom graph in Fig. 8 gives an example of querying the topic “Epson”, where we can observe some information on the topic, including posters, release time and post content.

As we have mentioned in the paper body, our model accurately forecasts users’ liking score for certain programs. With this valuable information, we can recommend these movies to the people whose liking score exceeds a certain threshold, which will increase the programs’ viewership and help businesses make better screening policies. For example, based on the users’ interest predicted by our model stated above, we can provide a TV series “Skip beat” to some target users, such as “The vagrant 1885838067”. The accurate recommendation not only improves users’ satisfaction, but also increases the viewership of “Skip beat”, meanwhile, it helps merchants get more benefits. Besides, it is useful for corporations to make effective marketing decisions, for example, proper advertisements may be shown to a given user and businesses can enhance their services to satisfy customers.

5. Conclusions and future work

Micro-blogging is one of the most popular social media platforms where the convenience, high update frequency and rich information have attracted millions of active users to join in. Users can publish posts on their daily lives and some trending topics. Trending topics are featured prominently to provide users with an up-to-date glimpse of what is happening in the real world and clearly reflect users’ interests. Users enthusiastically follow other users they are interested in to get relevant information of their interest.

In this paper, we propose a novel model to predict users’ interests in micro-blogging to help micro-blogging systems provide users better personalized information and advertising services. Our model is a probabilistic matrix factorization based approach. We present a user interest prediction framework fusing the social network structure and the evolution of users’ interest. In micro-blogging, the rich information and frequent updates make users’ interests more extensive and changeable over time, and then make users’ latent feature space and topics’ latent feature space change over time. Therefore, our model uses exponential decay function to obtain the mean matrix of user-latent feature matrix and the mean matrix of topic-latent feature matrix. The experimental results on Sina-weibo, one of the most popular micro-blogging sites in China, demonstrate that our model can improve the accuracy of predictions of users’ interest.

There are several directions future research can take. Firstly, there are some parameters in our proposed model, and different values of those parameters should affect the performance. Therefore, we would like to provide an efficient procedure for tuning the parameters automatically, such as Markov Chain Monte Carlo (MCMC) algorithm [13]. Secondly, we may unearth other effective factors to enhance the performance of the proposed model; for example, the retweeting relationships and discussions among users.
Acknowledgments

This research is supported by the NNSFC grants (no. 61172106, no. 71090402, no. 71002064) and the BJNSF grant (no. 4112062).

References


Hongyun Bao is a Ph.D. candidate in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. She received her B.S. degree in School of Mathematical Sciences from Capital Normal University, China, in 2008. Her research interests include information retrieval and web/text mining.

Qiudan Li (corresponding author) is an Associate Professor in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. She received a Ph.D. in Computer Science from Da Lian University of Technology, China, in 2004. Her research interests include web mining and mobile commerce applications. Her articles are published in Communications of the AIS, Decision Support Systems, Journal of the American Society for Information Science and Technology, IEEE Transactions on SMC, and Expert Systems with Applications.

Stephen Shaoyi Liao is a Professor at the Department of Information Systems and director of Advanced Transportation Information Systems Research Center, City University of Hong Kong. He is also a Visiting Professor and a Ph.D. Supervisor in USTC and Southwest Jiaotong University. He received a bachelor’s degree from Beijing University and a Ph.D. from the University of Aix-Marseille III and Institute of France Telecom. His research focuses on use of IT in e-business systems and transportation systems. His articles have been published in MISQ, Decision Support Systems, IEEE Transactions, Communications of the ACM, Information Science, Computer Software and other SCI journals.

Shuangyong Song is a Ph.D. candidate in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He received his B.E. degree in Biomedical Engineering from Beijing Jiaotong University, China, in 2007. His research interests include information retrieval and web/text mining.

Heng Gao is a Master Candidate in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He received his B.E. degree in Computer Science from China University of Mining and Technology, China, in 2010. His research interests include information retrieval, web/text mining and community question answering.