A Temporal-Topic Model for Friend Recommendations in Chinese Microblogging Systems

Nan Zheng, Shuangyong Song, and Hongyun Bao

***Abstract*—Due to its brief form and growing popularity, microblogging is becoming people's favorite choice for seeking information and expressing opinions. Messages received by a user mainly depend on whom the user follows. Thus, recommending users with similar interests may improve the experience quality for information receiving. Since messages posted by microblogging users reflect their interests, and the keywords in the messages indicate their main focus to a large extent, we can discover users' preferences by analyzing the user-generated contents. Moreover, users' interests are not static, on the contrary, they change as time goes by. Based on such intuitions, in this paper, we propose a temporal-topic model to analyze users' possible behaviors and predict their potential friends in microblogging. The model learns users' latent preferences by extracting keywords on aggregated messages over a period of time via a topic model, and then the impact of time is considered to deal with interest drifts. The experimental results of friend recommendations on Sina Weibo, one of the most popular microblogging sites in China, have demonstrated the effectiveness of our model.**

***Index Terms*—Friend recommendations, interest drifts, microblogging, Sina Weibo, topic model.**

I. Introduction

With the popularity of Web 2.0 technologies, microblogging is becoming one of the most prevalent social media platforms for Internet users. Sina Weibo, one of the most popular Twitter-like microblogging sites in China, had attracted more than 540 million registered users by December 4, 2013.[1] As a requirement of emerging and real-time information, microblogging is becoming people's favorite choice for seeking information and expressing opinions. Messages received by a user mainly depend on whom the user follows. Thus, to recommend users with similar interests may improve users' experience for information they desire to acquire. Users usually post microblogs to record daily life and express opinions. Therefore, posts published by users, to some extent, reflect their interests. By mining users' social behaviors and dynamics, we may help them find friends with similar interests, which may improve the users' experience, social interactions, and gain more business value for corporations [1]–[3]. For example, commercial advertisements may push their potential customers to read them. Also, recommending friends may enhance the social experience the site is designed to provide, which is an important indicator for the popularity of social media [4].

On Sina Weibo, users are allowed to post short messages with a limit of 140 characters each. Keywords in the message are good indications of users' interests. Thus, users' interests can be discovered by analyzing their keywords usage patterns. However, only keywords are not sufficient. If the words "movie" and "landscape" are the most frequent words used by user u_1, and the words "film" and "nature" often appear in user u_2's microblogs, it is clear that their common interests are movie and nature, although they use different descriptive word. By identifying the topics of words, we may easily find such common interests. Probabilistic topic models have been proved to be the powerful tools for identifying latent text patterns in the content. Latent Dirichlet allocation (LDA) achieves the capacity of generalizing the topic distributions so that the model can be used to generate unseen documents as well. LDA has also been applied to various works on Twitter [5], [6] to demonstrate its usefulness. Thus, we apply LDA to Sina Weibo microblogging for topic finding. Since our purpose is to discover the interests of each user rather than the topics of single messages, we aggregate multiple messages published by an individual user over a period of time into a single document. Then, we conduct the LDA model on keywords of the aggregated messages to discover users' preferences on Sina Weibo.

Users' interests are not static; contrarily, their interests may change as time goes by. Since the real-time and brevity features of microblogging lead to frequent updates of microblogs, users' interests are more extensive and changeable over time. Suppose that u_1 concerned "movie" two months ago, but now the main focus is on "landscape," while u_2 posted microblogs about nature two months ago, but now the posts are mainly about "film." Obviously, interest drifts exist in microblogging.

Manuscript received October 2, 2014; accepted December 7, 2014. This work was supported in part by the National Science Foundation of China under Project 71232006 and Project 61233001, in part by the Early Career Development Award of the State Key Laboratory of Management and Control for Complex Systems, and in part by the Youth Innovation Promotion Association of the Chine Academy of Sciences. This paper was recommended by Associate Editor L. Cao.

N. Zheng and H. Bao are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: nan.zheng@ia.ac.cn).

S. Song is with the Information Technology Laboratory, Fujitsu Research and Development Center Company, Ltd., Beijing 100025, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2015.2391262

[1] http://media.people.com.cn/n/2013/1204/c120837-23745866.html

Although u_1 and u_2 have common interests in movies and nature, as u_1's current interest is in nature, u_1 may continue to pay attention to a nature-related topic, while as u_2's current interest is in movies, u_2 may continue to focus on the movie area. In this case, it will be inappropriate to set the same weight on all the topics over time for each user to predict another's potential interests. In contrast, a higher weight should be assigned to more recent topics than those appearing a long time ago, since more recent preferences have greater influences in predicting users' potential interests than those earlier preferences. Therefore, taking temporal information into consideration may improve the accuracy of users' interest prediction.

In this paper, we propose a temporal-topic model to predict users' potential friends. The model first extracts users' topic distributions from keyword usage patterns of aggregated messages using a temporal approach. Then, it calculates user similarities over time based on users' topic distributions. Finally, users' potential interests on others are predicted according to user similarities over different periods of time via a temporal function. Based on the model, we conduct friend recommendations according to the predicted scores. To evaluate the effectiveness of the proposed method, we perform experiments on a real-world dataset crawled from Sina Weibo. Experimental results verify the usefulness of the proposed model and the importance of considering interest drifts when predicting users' potential interests.

This paper makes the following contributions: 1) the proposed model predicts users' potential interests using temporal latent semantic analysis; 2) we enhance the LDA model using time interval partition, which reveals users' interest drifts; and 3) the experimental results on real data from Sina Weibo show that users' interest drift follows a weekly changeable pattern. The remainder of this paper is organized as follows. We begin with a discussion of related work in the areas of discovering users' interests in microblogging and time-aware recommendations in Section II. Then, the proposed framework is described in Section III. A detailed description of our temporal-topic friend recommendation model is introduced in Section IV. Section V describes an empirical study as well as evaluation results and discussions. Finally, we conclude this paper in Section VI.

II. Related Work

The persistent popularity of microblogging systems has attracted many researchers' attention. Early researchers mainly focused on the characteristics of microblogging [6]–[8] and social network analysis [9], [10]. Recently, there has been an increasing interest in the field of information retrieval, such as event detection and tracking [11], [12], identification of influential people [13], [14], sentiment analysis [15]–[18], and personalized recommendations [19], [20]. This paper focuses on users' potential interests prediction, so in this section, we introduce the related work on user's interest analysis in microblogging systems and time-aware recommendations.

For user modeling, keyword extraction, semantic enrichment, and topic model are the three fundamental methods most widely used. Previous research simply built a bag-of-words profile based on user's microblogs [21], which may suffer from noisy and large amount of unrelated words. Liu *et al.* [22] mined the interests of Sina Weibo users via keyword extraction. Specially, they combined a translation-based method with a frequency-based method to extract appropriate interest keywords. Abel *et al.* [23] linked Twitter posts with related news articles from the web to create semantic user profiles. Xu *et al.* [24] modeled user posting behavior in Twitter by considering three factors: breaking news, posts from social friends and user's intrinsic interest via an extension of author-topic (AT) model, where user's intrinsic interest was represented by a distribution over latent topics. Hong and Davison [5] conducted standard LDA and extended AT model to predict popular Twitter messages and classified Twitter users and messages into categories. Their experiments showed that extended AT model did not yield better performance than LDA model. By analyzing different aggregation strategies of the data, the authors demonstrated that topic models learned from aggregated messages by the same user could lead to superior performance. Their work inspires us to emphasize users' aggregation messages on representing users' preferences. The crucial difference between this paper and the aforementioned studies is that we not only focus on detecting users' current preferences, but also attempt to predict users' potential interests in the near future. Instead of mining users' static historical behaviors, we aim to investigate users' interest drifts based on a sequence of snapshots of users' behaviors over a given time horizon. Therefore, we will introduce the related work on time-aware recommendations in the following.

Since more effective personalized recommendations depend on more accurate user's preference discovery, several works began to pay attention to the dynamic user interests. For instance, Ding and Li [25] presented a time weight item-based collaborative filtering via exponential time decay function to compute time weights for different items according to each user and each cluster of items. Xiong *et al.* [26] modeled time stamped user-item ratings as a 3-D tensor by assuming that each time feature vector depends only on its immediate predecessor. Xiang *et al.* [27] argued that user preferences often exhibit long-term and short-term factors, and then they proposed a session-based temporal graph model to capture users' dynamic preferences over time. Koenigstein *et al.* [28] presented a matrix factorization model incorporating temporal analysis of user ratings and item popularity trends to provide music recommendations. Zhang *et al.* [29] proposed an evolutionary topic pattern mining approach to discover changing of topic structures on a community question answering platform. The approach first extracted question topics via LDA in each time window, then discovered topic transitions based on cosine similarity, and finally analyzed life cycles of the extracted topics. Rafeh and Bahrehmand [30] proposed an adaptive collaborative filtering algorithm which takes time into account to reflect fluctuations in users' behavior over time. Liu *et al.* [31] developed a social temporal collaborative ranking model to recommend movies. To support time awareness, the authors used an expressive sequential matrix factorization model and a temporal smoothness regularization function to

TABLE I
LIST OF KEY NOTATIONS

SYMBOL	Description
N_u	Total number of users
N_w	Total number of keywords
$N_w^u(t)$	Total number of keywords of user u at time t
T	Number of topics
n	Total number of time intervals
I	Number of iterations in LDA model
α^t	Dirichlet prior for users at time t
β^t	Dirichlet prior for hidden topics at time t
γ	Kernel parameter in the exponential decay function
δ	Size of time interval
w_i^t	The unique word associated with the i-th token of user u at time t
z_i^t	The topic associated with w_i^t
$\theta_u(t)$	The multinomial distribution of topics specific to user u at time t
$\phi_z(t)$	The multinomial distribution of words specific to topic z at time t
S_t	Users' topical similarity matrix at time t
M_t	Users' keyword matrix at time t

tackle overfitting. Zheng *et al.* [32] conducted a brief survey of the dynamics models on social balance which examined the signs of links of the given networks at each time step during the whole process. Recently, Eirinaki *et al.* [33] proposed a trust-aware system for user recommendations which analyzed the semantics and dynamics of the implicit and explicit connections between users via a discounting factor. In [34], we applied an adaptive exponential forgetting function to imitate user's interest changes over time where each user was measured by his half-life behavior. All the above studies have demonstrated the importance of temporal information when discovering users' interests, but little has been done on recommendations in microblogging based on temporal information. To cope with trending topic recommendations in microblogging system, in [19], we proposed a probability matrix factorization based on the evolution of users' interest. An exponential decay function was adopted to obtain the mean matrix of user-latent feature matrix and the mean matrix of topic-latent feature matrix. This paper borrows the ideas of incorporating time factor with topic model for predicting users' potential interests in the Chinese microblogging system.

III. FRAMEWORK FOR FRIEND RECOMMENDATIONS

This section introduces the framework for the temporal-topic friend recommendation model. A glossary of notations used is given in Table I.

Fig. 1 shows the flowchart of our method. It first crawls users' microblogs over a period of time and puts them in the microblog corpus. After preprocessing, Chinese word segment and part-of-speech (POS) tagging are conducted, and keywords are extracted based on the POS tagging. Then, the collected microblogs are split into time intervals. With this temporal information, users' temporal interests are discovered based on the temporal-topic model. The model first learns users' topics in each time interval, and then calculates user similarities based on temporal topic distributions. In addition, a temporal function is used to compute the changing of interests. Finally, users' potential interests in others can be predicted based on the sequence of users' interests along a timeline. Below, we discuss each step of the framework in detail.

IV. TEMPORAL-TOPIC FRIEND RECOMMENDATIONS

We introduce our temporal-topic friend recommendation model step-by-step in this section.

A. Preprocessing

In Sina Weibo, if a user reposts others' messages without any comments, the system will add "forwarding microblogs" automatically. Such a denotation does not have any effect on users' interests; therefore, we remove it from messages, but keep the content of the reposted messages, since reposts represent users' interests on the related content. In addition, we remove URLs and other nontexts from microblogs.

B. POS and Keyword Extraction

Since this paper is carried out on Chinese microblogs, we face a problem caused by Chinese word segmentation. To address this problem, we perform word segmentation and POS tagging for messages. We apply SWJTU Yebol Chinese word segmentation platform[2] to preprocess the corpus. SWJTU Yebol Chinese word segmentation platform proposes a new Chinese word segmentation approach based on integration of human intelligence, big data, and machine learning. Based on POS tagging, we extract nouns, abbreviations, idioms, and academic vocabularies as meaningful notional words which form keywords for further analysis.

C. Time Interval Partition

Users' interests change as time goes by, which reveals in Sina Weibo that users' microblogs may focus on different topics at different periods of time. Therefore, users' dynamically changing interests can be expressed as a sequence of keyword collections in microblogs at different time intervals, i.e., $\boldsymbol{M} = M_1 \cup M_2, \ldots, \cup M_n$. Each M_t denotes a temporal user-keyword matrix at the tth time interval, where $M_t \in R^{Nu \times Nw}$, N_u and N_w are the numbers of users and keywords, respectively. Each row of M_t contains the word counts at the tth time interval for a particular user, whereas each column of M_t contains the counts by different users for a certain word at the tth time interval.

D. Topic Finding

It has been described in Section I that only keywords are not sufficient for discovering users' interests. As the existence

[2]http://ics.swjtu.edu.cn/

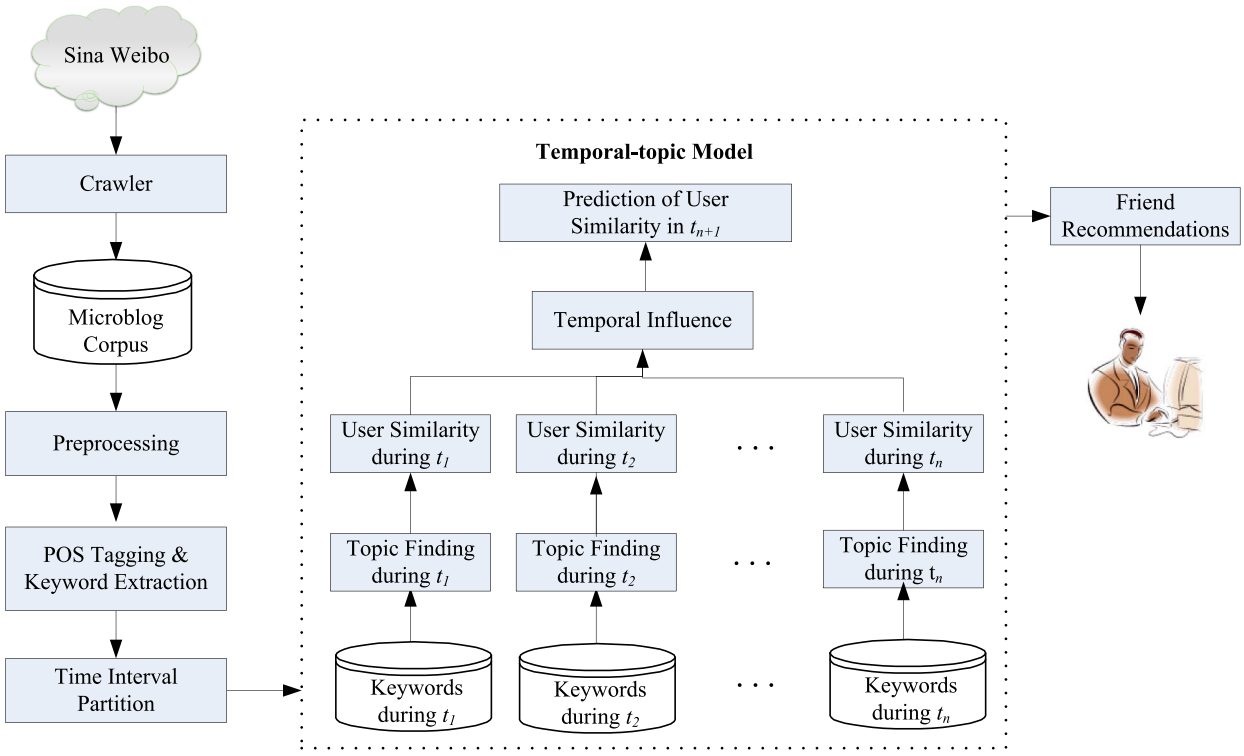

Fig. 1. Overview of the research design framework.

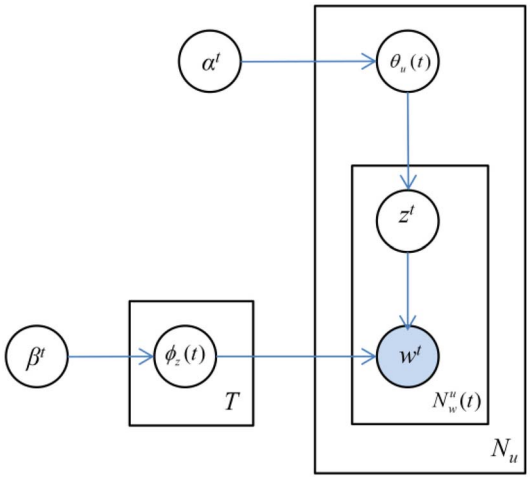

Fig. 2. Graphical representation of LDA model using plate notation.

Algorithm 1 LDA model [35]

For each topic $z^t = 1, \ldots, T$:
  Draw a multinomial distribution over keywords, $\phi_z(t)$, from Dirichlet prior β^t;
For each user $u = 1, \ldots, N_u$:
  Draw a multinomial distribution over topics, $\theta_u(t)$, from Dirichlet prior α^t;
  For each word token $i = 1, \ldots, N_w^u(t)$ used by user u:
    Sample a topic ${z_i}^t$ from multinomial distribution $\theta_u(t)$ conditioned on u_i;
    Sample a word ${w_i}^t$ from multinomial topic distribution $\phi_z(t)$ conditioned on ${z_i}^t$.

of synonymy, it needs to find the hidden topics from the keyword usage patterns. Since the goal is to find topics that each microblogging user is interested in rather than topics that each microblog is about, we treat the microblogs published by an individual user at the tth time interval as a big document. Then, each row of sub-collection M_t is treated as a bag-of-words document which essentially corresponds to a user. To find user temporal topics in M_t, or to find temporal topics of each document in M_t, we apply the LDA model [35].

Each user is associated with a mixture of different topics, and each topic is represented by a probabilistic distribution over keywords. Formally, each of a collection of N_u users is associated with a multinomial distribution over T topics, which is denoted as $\theta_u(t)$ at time t. Each topic is associated with a multinomial distribution over keywords, denoted as $\phi_z(t)$. $\theta_u(t)$ and $\phi_z(t)$ have Dirichlet prior with hyper-parameters α^t and β^t, respectively. For each keyword of user u, a topic z^t is sampled from the multinomial distribution $\theta_u(t)$ associated with user u at time t, and a keyword w^t from the multinomial distribution $\phi_z(t)$ associated with topic z^t is sampled consequently. This generative process is repeated $N_w^u(t)$ times to form user u's collection of keywords. The generative process can be graphically represented using commonly-used plate notations in Fig. 2. In the graphical notation, the shaded and unshaded variables indicate observed and latent variables, respectively. Arrows indicate conditional dependencies between variables while plates refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples.

Let ${z_1}^t, \ldots, {z_T}^t$ be T topics to be extracted at time t; the generative process of extracting temporal topics from M_t is given in Algorithm 1.

The user-topic distribution at tth time interval $\theta(t)$ and the topic-word distribution at tth time interval $\phi(t)$ are two main variables of interest. They are estimated using the Gibbs sampling method [36], a Markov chain Monte Carlo algorithm to sample from the posterior distribution over parameters

$$\phi(t)_{mj} = \frac{C^{\mathrm{WT}}_{mj} + \beta^t}{\sum_{m'=1}^{N^u_w(t)} C^{\mathrm{WT}}_{m'j} + N^u_w(t)\beta^t} \tag{1}$$

$$\theta(t)_{hj} = \frac{C^{\mathrm{UT}}_{kj} + \alpha^t}{\sum_{j'=1}^{T} C^{\mathrm{UT}}_{hj'} + T\alpha^t} \tag{2}$$

where C^{WT} is the word-topic matrix of counts with dimensions $N_w \times T$, and C^{WT}_{mj} is the number of times for word m to be assigned to topic j, excluding the current instance. Similarly, C^{UT} is the user-topic matrix of counts with dimensions $N_u \times T$, where C^{UT}_{hj} is the number of times user h is assigned to topic j, excluding the current instance. By applying the LDA model to the entire collection $\boldsymbol{M}$, we can obtain n temporal topics for each user.

E. User Similarity Calculation

After row normalizing $\theta(t)$ to $\theta'(t)$, the ith row of matrix $\theta'(t)$ provides an additive linear combination of factors to indicate user i's interests over T topics at the tth time interval. The higher weight user i is assigned to a factor, the more interest user i has in the relevant topic. It has been demonstrated in [13] that microblogger follows a friend because he is interested in some topics the friend is publishing. Therefore, for friend recommendations, we aim to find users' topic similarity based on the normalized user-topic distribution $\theta'(t)$. As Jensen–Shannon divergence is a symmetric method of measuring the difference between two probability distributions, we use it to denote the topical difference between two probability distributions $\theta'(t)_{i\cdot}$ and $\theta'(t)_{j\cdot}$.

$$D_{JS}(i,j)^t = \frac{1}{2}\left(D_{\mathrm{KL}}\left(\theta'(t)_{i\cdot} \| H(t)\right) + D_{\mathrm{KL}}\left(\theta'(t)_{j\cdot} \| H(t)\right)\right) \tag{3}$$

where $H(t)$ is the average of the two probability distributions, $H(t) = 1/2(\theta'(t)_{i\cdot} + \theta'(t)_{j\cdot})$. D_{KL} is the Kullback–Leibler (KL) divergence between topic distributions

$$D_{\mathrm{KL}}\left(\theta'(t)_{i\cdot} \| H(t)\right) = \sum_k \theta'(t)_{ik} \log \frac{\theta'(t)_{ik}}{H(t)_k}. \tag{4}$$

Since Jensen–Shannon divergence does not obey the triangle inequality, it cannot be proved that the shortest path between two points in Euclidean geometry is a straight line. However, its square root has been proved to comply with the triangle inequality [37]. Therefore, we use the square root as a metric for topical difference to calculate topical similarity, indicating that direct user similarity impact is more important than indirect user similarity impact. Topic similarity between user i and user j is calculated as follows [13]:

$$\mathrm{sim}(i,j)^t = 1 - \mathrm{dist}(i,j)^t = 1 - \sqrt{2{*}D_{JS}(i,j)^t}. \tag{5}$$

After calculating topical similarities over all users in n time intervals, we get n temporal similarity matrices $S_1, S_2, \ldots, S_n$, where $S_t \in R^{Nu \times Nu}$.

F. Temporal Influence

In this step, we desire to utilize users' sequential topical similarity matrices $\{S_1, S_2, \ldots, S_n\}$ to predict users' potential interests in the near future. Generally speaking, users' historical favorites may influence his future interests, and more recent interests may have stronger impact on the future preference prediction than earlier interests. To imitate the influence of historical behaviors, we apply the exponential decay function [19], which has been proved to be an effective function to measure interest drifts

$$f(t) = \exp\left(-\frac{n-t}{\gamma}\right) \quad (t \in \{1, 2, \ldots, n-1\}, \gamma > 0) \tag{6}$$

where γ is the kernel parameter, and the value of $n - t$ represents the time interval between the tth time instant and the current time n. For an earlier time $t, n-t$ gets a higher value, which results in a smaller influencing value of $f(t)$. On the contrary, a more recent time t can result in a higher influencing value of $f(t)$. Obviously, the exponential decay function can gradually discount the history of past behavior.

G. Friend Recommendation

Finally, we utilize the exponential decay function with kernel parameter γ to predict users' potential interests on others at time t as follows:

$$P_n = \sum_{t=1}^{n-1} \exp\left(-\frac{n-t}{\gamma}\right) S_t \tag{7}$$

where P_n is the probability matrix of potential interests among users at time n. A higher score indicates that the two users have greater similarity and that they may have greater likelihood of becoming a friend. Finally, users are sorted by the score and those with higher scores are recommended to the target user.

H. Complexity Analysis

The main computation of our model involves three parts.

1) Suppose that the process of LDA model needs I iterations to reach convergence. Then, the cost of conducting topic finding in each time interval is $O((N_w + N_u)^* T^* I)$.
2) In each time interval, $O(N^*_u(N_u - 1))$ operations are required to compute user similarity.
3) Finally, the cost of calculating friend recommendations is $O(N^*_u n)$.

Thus, the total complexity of computing the temporal-topic friend recommendation model is $O((N_w + N_u)^* T^* I^* n + N^{2*}_u n) = O(N^*_w T^* I^* n + N^{2*}_u n)$. Obviously, the complexity of our model is proportional to the keyword size and the square of the user size. Since LDA has the ability of generating unseen documents, when new messages appear in users' microblogs, it may only calculate on the updated data. Thus, the proposed algorithm is efficient and feasible for online recommendation processing.

V. Experiments

A. Dataset

To examine the performance of our model, we conducted an experimental study on real-world datasets. We randomly

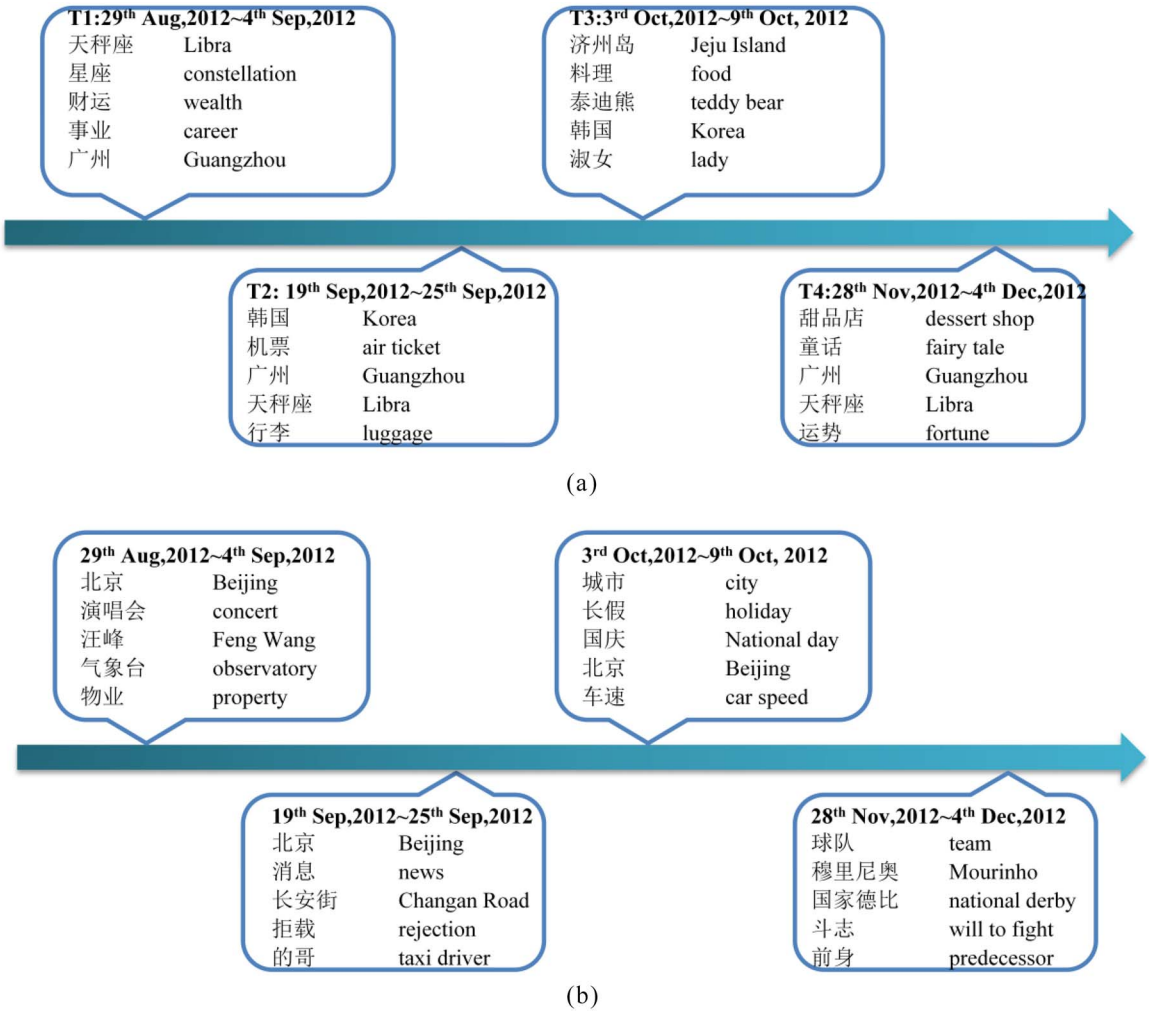

Fig. 3. Lists of top-ranked keywords during different time period for (a) User #1 and (b) User #2.

selected 20 users as the seed then crawled by their latest 50 friends through Sina Weibo API. For each user, we collected his full set of microblogs between 29th August 2012 and 4th December 2012 (98 days). Each microblog includes its author, content, the time it was posted, author's friend count, follower count, etc. After removing the users who published less than 300 posts and whose friends and followers were less than 50 or more than 3000, the dataset contained 1233 users. We then crawled all the friends of each individual user along with the timestamp over the same time period, resulting in 193 762 distinct users. These 193 762 users were then pruned off to the 1233 users belonging to our dataset. Finally, our dataset contained 1233 users with 347 284 microblogs and the friendship density was 2.67%. In our experiments, we set the hyper-parameters $\alpha^t = 50/\mathrm{T}$ and $\beta^t = 0.01$, as widely adopted in [36]. Gibbs sampling was repeated for $I = 50$ iterations.

B. Evaluation Metric

Since users are usually more concerned with top-ranked recommended users, top-ranked users in the results should be rewarded more heavily than those ranked lower. The metric of mean average precision (MAP) [38] is therefore adopted, which was initially devised to measure retrieval results. In our experiments, MAP was defined as the average precision at each point a real friend is recommended

$$\mathrm{MAP} = \frac{\sum_{i=1}^{n^+} \frac{\sum_{j=1}^{i} r(j)}{i} * r(i)}{\sum_{i=1}^{n^+} r(i)} \tag{8}$$

where i is the position of the friend in the recommendation list and n^+ is the number of real friends appeared in the recommendation list. $r(i)$ is a binary value: when the suggested friend at position i is the user's real friend, $r(i)$ is set to 1, or 0 otherwise. The mean of the MAPs of all test users is referred to as mMAP

$$\mathrm{mMAP} = \frac{\sum_{i=1}^{N_u} \mathrm{MAP}(i)}{N_u} \tag{9}$$

where MAP(i) represents the MAP value for the ith user.

C. Results and Discussions

1) Interest Drifts: Before performing friend recommendations, we first discuss users' interests change with time in Sina Weibo. Fig. 3 illustrates interests of two users at different time periods, with each time interval represented by a list of top-ranked keywords. As shown, User #1 was concerned

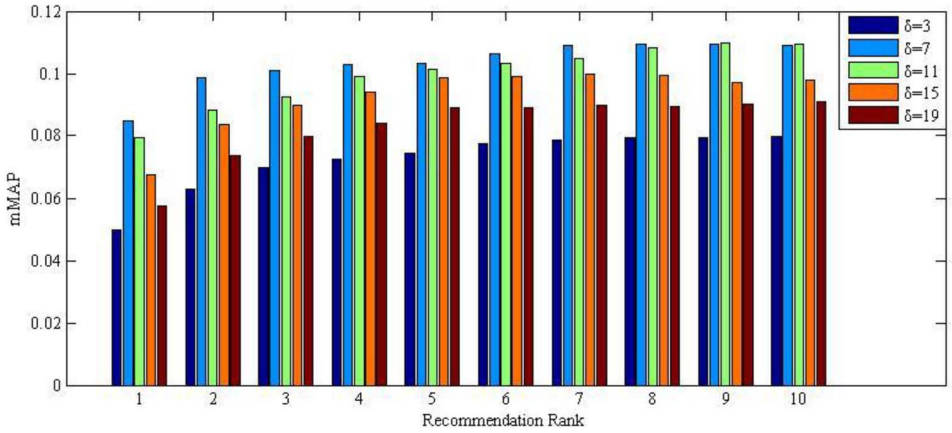

Fig. 4. mMAP results of different values of δ for top-k ($k = 1, 2, \ldots, 10$) friend recommendations.

with constellation in T1, shifted to air ticket to Korea in T2, then to travel in Jeju Island in T3, and finally to desserts in T4. Although User #1 had different interests for different time periods, there were relations between them. For example, he was first concerned with an air ticket to Korea in T2 and then had a visit in Korea in T3. Also, he paid attention to constellation during almost all the time, although with different degrees. For User #2, he began with Beijing concert of singer F. Wang in T1, Beijing news about taxi drivers in T2, traffic in Beijing at national day in T3 and evolved into soccer in T4. Obviously, interest drifts exist in Sina Weibo users, and their future potential interests are influenced by their past behaviors.

2) Parameter Setting: The time-interval length δ and the kernel parameter γ in (6) are two key parameters in our model that determine the performance of the prediction. The time-interval length δ should be long enough to collect sufficient microblogs of users and yet be short enough to capture the evolution of users' interests. We enumerate a list of values for δ and pick up the value with which the algorithm performs best. Different δ will divide the data of 98 days into $n = \lfloor 98/\delta \rfloor$ time intervals. The data in the nth interval is regarded as the ground-truth, and historical data is used for prediction. Fig. 4 shows the results of top-k ($k = 1, 2, \ldots, 10$) friend recommendations for each user. As shown, when $\delta = 7$ the model gets the best performance, so we set $\delta = 7$ in our following empirical work. In this way, our final investigation is performed on a sequence of weekly time intervals. This gives us 14 time intervals in total. The former 13 weeks (from 29th August 2012 to 27th November 2012) are treated as background dataset, and the objective is to predict users' potential friends in the last time interval (from 28th November 2012 to 4th December 2012).

The kernel parameter γ is used to adjust the impact of historical behaviors on predicting users' potential friends. A smaller γ indicates that an earlier behavior plays a smaller role on users' future interests than a recent behavior. In this paper, we enumerate a list of values for different γ and pick up the value with which the algorithm performs best. Fig. 5 shows the top-ten friend recommendation results when varying γ from 1 to 19 with an interval of 2. As shown, the curve peaks at $\gamma = 11$, so we set $\gamma = 11$ in the following empirical work.

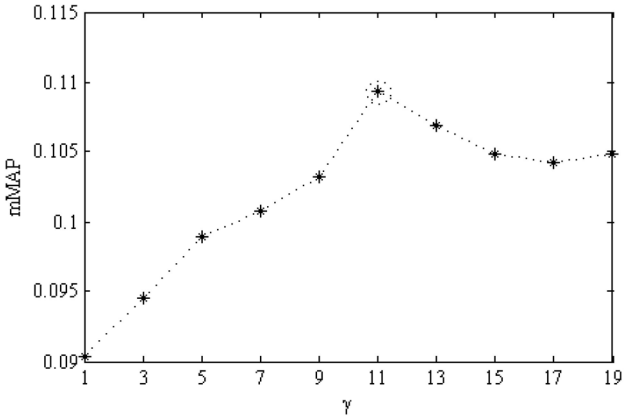

Fig. 5. mMAP results of different values of γ for top ten friend recommendations.

3) Evaluation of Friend Recommendations: In this section, to evaluate the performance of the proposed algorithm, we compare it with the other four models.

a) *LDA:* This model discovers users' topic distributions according to their keyword usage patterns, and then calculates user similarity based on the method described in Section IV-E. The difference between LDA and our algorithm is that in LDA the temporal information is not taken into account. It just uses the former 13 weeks' data to build the model to predict users' future interests.

b) *In-Degree:* Since users seem to follow influential people on the microblogging platform [13], it is reasonable to add such people as users' friends. The in-degree method calculates the influence of users by the number of followers and then suggests friends based on users' influence.

c) *User-Based [39]:* This method is based on the intuition that users who have more common friends may be more similar to each other. Therefore, friends can be suggested according to the number of users' common friends. This is the measurement currently employed by Sina Weibo and other social network services, such as *renren.com*. Since the user-friend matrix is sparse, we use Jaccard coefficient to measure user similarity.

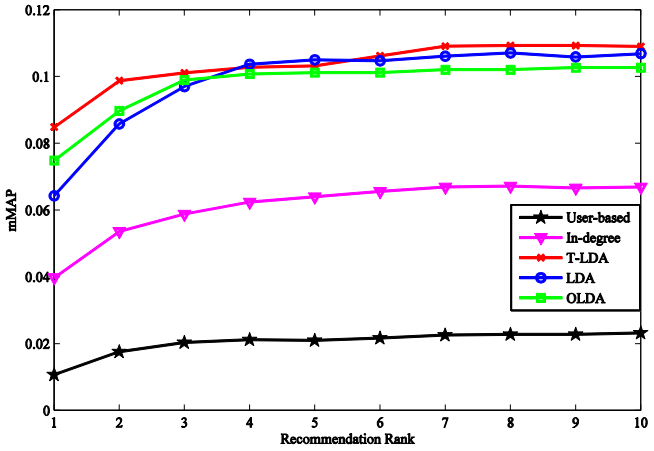

Fig. 6. mMAP results of top-k ($k = 1, 2, \ldots, 10$) friend recommendations.

d) *On-Line LDA (OLDA) [40]:* This model also divides the corpus into time slots to discover topics. Different from our method, OLDA applies LDA while using the topic models learned from previous time slices as a prior for the current model. To calculate user similarity, we follow the same process described in Section IV-E. Finally, friends are recommended according to users' topic similarity.

Each of the model is performed to predict top-k (k = 1, 2, …, 10) friends for each user according to the former 13 weeks' data and the ground-truth is the actual friends generated by each user in the 14th week. For convenience of presentation, the proposed method is denoted as T-LDA.

The evaluation results are shown in Fig. 6 in terms of mMAP. It is observed that LDA-based methods outperform in-degree and user-based methods, suggesting that the main reason for adding friends in Sina Weibo is the topical similarity between users based on our dataset. Among the five methods, the user-based model performs the worst. On the one hand, it ignores users' topical similarity and their interest drifts, which are good indications of users' preferences. On the other hand, the sparsity of user-friend matrix results in a low coverage of the user-based model, leading to the low performance of friend recommendations. As shown, the in-degree model achieves better performance than the user-based approach. This may be due to the fact that one of the purposes of participation in microblog is to seek information. As influential users may more likely post high-quality microblogs, users tend to follow them to improve experience quality for information receiving.

Compared with the three LDA-based models, T-LDA performs better than LDA and OLDA, with exceptions at $k = 4$ and $k = 5$. Especially, the great improvements at $k = 1$ and $k = 2$ of T-LDA and OLDA, as compared with LDA indicate that adding temporal information can improve the quality of recommendations for the top ranks. Adding temporal influence may distinguish users' current interests from the interests that were a long time ago. More recent interests may play a more important role in predicting users' future interests. Thus, T-LDA and OLDA achieve better performances for friend recommendations at top ranks. The comparison between T-LDA and OLDA shows the benefit of exponential temporal decay influence when imitating interest drifts. As in OLDA model, the historical data is tracked as prior patterns, the influence of long-term interests may decline. In our T-LDA model, the exponential decay function gradually discounts the history of past behavior, which highlights current interests as well as considers long-term interests. The improvement achieved by our T-LDA algorithm confirms that topical similarity is good indication of users' preferences and adding interest drifts results in more accurate predictions. Overall, the low mMAP values are mainly attributed to the low friendship density of our dataset, which is only 2.67%.

VI. Conclusion

In this paper, we propose a temporal-topic model for friend recommendations in Chinese microblogging systems. The model first discovers users' latent preferences during different time intervals based on keywords extracted from the aggregated microblogs through a topic model. Then, it calculates user similarities in each time interval based on temporal topic distributions. After that, an exponential decay function is used to measure interest drifts. Finally, users' potential interests on others can be predicted based on the sequence of users' interests along the timeline. Based on the model, we conducted friend recommendations and the experimental results showed that our model is effective. For future work, we plan to conduct our experiments on users who have less friends and followers to show if our model is useful for the cold-start problem of personalized recommendations. We also aim to unearth other factors to enhance the performance of the proposed model, such as social relationships among users (i.e., followers, followees), the sentiment of microblogs, users' location information, etc. We also plan to investigate other state-of-the-art models with temporal evolvement and compare the performances of different methods on friend recommendations. Other datasets such as Twitter will be tested for the usefulness and effectiveness of the model.

Acknowledgment

The authors would like to thank Prof. F.-Y. Wang for his instructive advice and valuable suggestions and in completion of this paper. They also would like to thank the anonymous reviewers for providing valuable comments. By responding to those significant comments, they were able to improve the content and presentation of this paper.

References

[1] F.-Y. Wang, "Toward a paradigm shift in social computing: The ACP approach," *IEEE Intell. Syst.*, vol. 22, no. 5, pp. 65–67, Sep./Oct. 2007.

[2] F.-Y. Wang, K. M. Carley, D. Zeng, and W. Mao, "Social computing: From social informatics to social intelligence," *IEEE Intell. Syst.,* vol. 22, no. 2, pp. 79–83, Mar./Apr. 2007.

[3] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Inf. Sci.*, vol. 275, pp. 314–347, Aug. 2014.

[4] M. Moricz, Y. Dosbayev, and M. Berlyant, "PYMK: Friend recommendation at myspace," in *Proc. ACM SIGMOD Int. Conf. Manage. Data.,* Indianapolis, IN, USA, 2010, pp. 999–1002.

[5] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proc. 1st Workshop Soc. Media Anal.*, Washington, DC, USA, 2010, pp. 80–88.

[6] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in *Proc. Int. AAAI Conf. Weblogs Soc. Media*, Menlo Park, CA, USA, 2010, pp. 130–137.

[7] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 591–600.

[8] D. Zhao and M. B. Rosson, "How and why people Twitter: The role that micro-blogging plays in informal communication at work," in *Proc. ACM Int. Conf. Support. Group Work*, Sanibel Island, FL, USA, 2009, pp. 243–252.

[9] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD 1st SNA-KDD Workshop Web Min. Soc. Netw. Anal.*, San Jose, CA, USA, 2007, pp. 56–65.

[10] W. X. Zhao *et al.,* "Comparing Twitter and traditional media using topic models," in *Advances in Information Retrieval*. Berlin, Germany: Springer, 2011, pp. 338–349.

[11] S. Song, Q. Li, and X. Zheng, "Detecting popular topics in microblogging based on a user interest-based model," in *Proc. Int. Joint Conf. IEEE Neural Netw. (IJCNN)*, Brisbane, QLD, Australia, 2012, pp. 1–8.

[12] S. Song, Q. Li, and H. Bao, "Detecting dynamic association among Twitter topics," in *Proc. 21st Int. Conf. Companion World Wide Web*, Lyon, France, 2012, pp. 605–606.

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Min.*, New York, NY, USA, 2010, pp. 261–270.

[14] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. 4th Int. AAAI Conf. Weblogs Soc. Media (ICWSM)*, Menlo Park, CA, USA, 2010, pp. 10–17.

[15] J. Lehmann, B. Goncalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in Twitter," in *Proc. 21st Int. Conf. World Wide Web*, Lyon, France, 2012, pp. 251–260.

[16] A. Agarwal, V. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proc. Workshop Lang. Soc. Media Assoc. Comput. Linguist.*, Stroudsburg, PA, USA, 2011, pp. 30–38.

[17] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. Lang. Resour. Eval. Conf. (LREC)*, Valletta, Malta, 2010, pp. 1320–1326.

[18] D. H. Yang and G. Yu, "A method of feature selection and sentiment similarity for Chinese micro-blogs," *J. Inf. Sci.,* vol. 39, no. 4, pp. 429–441, 2013.

[19] 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," *Decis. Support Syst.,* vol. 55, no. 3, pp. 698–709, 2013.

[20] K. Chen *et al.,* "Collaborative personalized tweet recommendation," in *Proc. 35th Int. ACM SIGIR Conf. Res Dev. Inf. Retrieval*, Portland, OR, USA, 2012, pp. 661–670.

[21] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: Experiments on recommending content from information streams," in *Proc. 28th Int. Conf. Human Factors Comput. Syst.*, Atlanta, GA, USA, 2010, pp. 1185–1194.

[22] Z. Liu, X. Chen, and M. Sun, "Mining the interests of Chinese microbloggers via keyword extraction," *Front. Comput. Sci.*, vol. 6, no. 1, pp. 76–87, 2012.

[23] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Semantic enrichment of Twitter posts for user profile construction on the social web," in *Proc. 8th Extended Semantic Web Conf. Semantic Web Res. Appl.*, Berlin, Germany, 2011, pp. 375–389.

[24] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, "Modeling user posting behavior on social media," in *Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval*, Portland, OR, USA, 2012, pp. 545–554.

[25] Y. Ding and X. Li, "Time weight collaborative filtering," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, Bremen, Germany, 2005, pp. 485–492.

[26] L. Xiong, X. Chen, T. Huang, J. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with Bayesian probabilistic tensor factorization," in *Proc. 10th SIAM Int. Conf. Data Min.*, Sydney, NSW, Australia, 2010, pp. 211–222.

[27] L. Xiang *et al.,* "Temporal recommendation on graphs via long- and short-term preference fusion," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, 2010, pp. 723–731.

[28] N. Koenigstein, G. Dror, and Y. Koren, "Yahoo! music recommendations: Modeling music ratings with temporal dynamics and item taxonomy," in *Proc. 5th ACM Conf. Recommender Syst.,* Vienna, Austria, 2011, pp. 165–172.

[29] Z. Zhang, Q. Li, and D. Zeng, "Mining evolutionary topic patterns in community question answering systems," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans,* vol. 41, no. 5, pp. 828–833, Sep. 2011.

[30] R. Rafeh and A. Bahrehmand, "An adaptive approach to dealing with unstable behaviour of users in collaborative filtering systems," *J. Inf. Sci.*, vol. 38, no. 3, pp. 205–221, 2012.

[31] N. N. Liu, L. He, and M. Zhao, "Social temporal collaborative ranking for context aware movie recommendation," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, 2013, Art. ID 15.

[32] X. Zheng, D. Zeng, and F.-Y. Wang, "Social balance in signed networks," in *Information Systems Frontiers*. New York, NY, USA: Springer, 2014, pp. 1–19.

[33] M. Eirinaki, M. D. Louta, and I. Varlamis, "A trust-aware system for personalized user recommendations in social networks," *IEEE Trans. Syst., Man, Cybern., Syst.,* vol. 44, no. 4, pp. 409–421, Apr. 2014.

[34] N. Zheng and Q. Li, "A recommender system based on tag and time information for social tagging systems," *Expert Syst. Appl.,* vol. 38, no. 4, pp. 4575–4587, 2011.

[35] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.,* vol. 3, pp. 993–1022, Jan. 2003.

[36] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook Latent Semantic Anal.,* vol. 427, no. 7, pp. 424–440, 2007.

[37] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1858–1860, Jul. 2003.

[38] K. Kishida, "Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments," Res. Center for Inf. Resources, Nat. Inst. Inf., Chiyoda, Japan, Tech. Rep. NII-2005-014E, 2005.

[39] J. Wu *et al.,* "Predicting quality of service for selection by neighborhood-based collaborative filtering," *IEEE Trans. Syst., Man, Cybern., Syst.,* vol. 43, no. 2, pp. 428–439, Mar. 2013.

[40] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Proc. 8th IEEE Int. Conf. Data Min.* (*ICDM*), Pisa, Italy, 2008, pp. 3–12.

Nan Zheng received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

She is an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. Her current research interests include personalized recommendation, web mining, and information retrieval.

Shuangyong Song received the Ph.D. degree from the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is a Researcher with Information Technology Laboratory, Fujitsu Research and Development Center Company, Ltd., Beijing. His current research interests include information retrieval, web/text mining, and natural language processing.

Hongyun Bao received the B.S. degree from the School of Mathematical Sciences, Capital Normal University, Beijing, China, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2008 and 2013, respectively.

She is a Research Assistant with the Institute of Automation, Chinese Academy of Sciences. Her current research interests include information retrieval, web/text mining, and knowledge engineering.