Comput Syst Sci & Eng (2018) 2: 95–103

International Journal of
Computer Systems
Science & Engineering

Research on Tourist Routes Recommendation Based on the User Preference Drifting Over Time

Chunjing Xiao[1]*, Yongwei Qiao[2], Kewen Xia[1] and Yuxiang Zhang[3]

[1] *School of Electronics & Information Engineering, Hebei University of Technology, School of Computer Science & Technology, Civil Aviation University of China, Tianjin 300401, China. E-mail: kwxia@hebut.edu.cn*
[2] *Engineering & Technical Training Center, Civil Aviation University of China, Tianjin 300300, China*
[3] *School of Computer Science & Technology, Civil Aviation University of China, Tianjin 300401, China*

Tourist routes recommendation is a way to improve the tourist experience and the efficiency of tourism companies. Session-based methods divide all users' interaction histories into the same number sessions with fixed time window and treat the user preference as time sequences. There have few or even no interaction in some sessions for some users because of the high sparsity and temporal characteristics of tourist data. That lead to many session-based methods can not be applied to routes recommendation due to aggravate the sparsity. In order to better adapt and apply the characteristics of tourism data and alleviate the sparsity, a tourist routes recommendation method based on the user preference drifting over time is proposed. Firstly, the sparsity, temporal context, tourist age and price characteristics of tourism data are analyzed on a real tourism data. Secondly, based on the results of analysis, tourist interaction history is dynamic divided into different number of sessions and the tourist's evolving profile is then constructed by mining his probabilistic topic distribution in each session using Latent Dirichlet Allocation (LDA) and the time penalty weights. Then, the tourist feature vector based on the tourist age, the price and season of his tourism is modeled and a set of nearest neighbors and the candidate routes is selected base on it. Finally, the routes are recommended according to the similarities of probabilistic topic distributions between the active tourist and routes. Experimental results show that the proposed method can not only effectively adapt to the characteristics of tourism data, but also improve the effect of recommendation

Keywords: Route Recommendation; LDA; User Preference; Time Penalty; Feature Vector

1. INTRODUCTION

With the improvement of people's living standards, tourism has become an important way of leisure and entertainment. According to the statistics in recent years, the number of tourists and tourism income are growth in the speed of more than 10%. In order to attract more tourists, tourism companies need to understand the needs of tourists and develop a variety of attractive routes. But it is difficult for tourists to choose their own routes from a large number of routes. Therefore, obtaining the user's travel needs to recommend the tourist routes has became a problem to be solved in the tourism industry.

Recommender systems [1] which are the main ways to solve the problem of "information overload" are applied to the routes recommendation to greatly enhance the experience of tourists and to bring benefits to tourism companies. It is influenced by tourist preference changing over time when tourists choose their routes. So modeling the tourist preference drift over time is essential to improve the recommendation results. LDA (Dirichlet Allocation Latent) is an important method for text mining to discover text topics, and has been extended and applied to the potential interest in the field of recommendation [2,3]. Session-based collaborative filtering divide all users' interaction histories into the same number different stages with fixed time window and the preference is represented by the temporal sequences [4,5,6,7,8], that can be used to improve the recommendation accuracy. Compared to the recommendation of products, movies, there will have few or even no interaction in some sessions for

*Corresponding Author. E-mail: chunjingxiao@163.com

some users if their histories are divided into the same number sessions with fixed time window because of the high sparsity, the seasonal and temporal of the tourism data. That will not only aggravate the sparsity, but also make many session-based methods unavailable directly to routes recommendation. Although many recommendation algorithms and systems for tourists have emerged, for example, He et. al. dug out the hidden theme from documents based on LDA and determined which route is most suitable to the user according to grades generated for each user to each route by using Collaborative Filtering algorithm [9]. Tourist route intelligent recommendation system based on Hadoop used distributed association rules calculation to solve the secure storage and the fast access of the large amounts of data [10], these works can not meet the unique characteristics of tourism data.

In this paper, in order to adapt to the high sparsity, seasonal and temporal characteristics and patterns of tourism data, and model the user preference drift over time to improve the accuracy of the tourism routes, a route recommendation model based on dynamic dividing a user's interaction history is proposed. Firstly, the high sparsity, temporal, price and age characteristics of tourism data are analyze. Secondly, every tourist history is divided into different stages based on the temporal characteristics of tourism data, and the tourist evolving preference is modeled by extracting the probabilistic topic model which represents the user latent interest in every stage using LDA and defining the time decreasing weight. Then, the tourist's feature vector is established according to the characteristics of the age of tourists, tourism season and price to obtain a set of neighbors and candidate routes for an active tourist. Finally, the routes are recommended based on the relevant model of the probabilistic topic distribution between candidate routes and an active tourist. A large number of experimental results on actual tourism data set show that the method can effectively use the characteristics of the tourism data, and recommend accurately the routes.

The major contributions of this paper are as follows:

- We demonstrate the high sparsity and temporal characteristics of tourism data. The analysis shows that it is unreasonable to divide tourism history into the same number stages with fixed time window which can worsen sparsity issue.
- we propose a novel method which dynamic divides every user's interaction history into different number stages based on the tourism data and the user's preference drift over time is modeled by using LDA for each session and a time sensitive weighting scheme to capture the user's evolving interest.
- We employ LDA as the language model to detect the probabilistic topic distribution for each tourist in every session which represents the latent factors affecting the tourist's choice on routes. It is easy to mining the preference changes for each tourist on topic and predict the tourist's preference trend on routes, that is important to build profile.
- We take temporal context, tourist age, route price and travel season into account when the neighbors are selected to recommend the routes to the active tourist, which is suitable for the characteristic of actual tourism data.
- We conduct a large set of experiments to evaluate the performance of our method and compare our method with other state-of-the-art methods.

The remainder of this paper is organized as follows. In section 2, we provide a brief overview of related works. Section 3 analyses the sparsity and temporal characteristics of tourism data and proposes a novel method to divide the tourism history into sessions and models the user's preference based on temporal domain division by using LDA and a time sensitive weighting scheme. The method of neighbor selection and routes recommendation is introduced in section 4. The results of an empirical analysis are presented in section 5, followed by a conclusion in section 6.

2. RELATED WORKS

In this section, we briefly review tourism recommendation systems, LDA-based recommendation and session-based temporal dynamic recommendation.

2.1 Tourism Recommendation

The whole process of tourism is covered by recommendation. Routes can be recommended before trip, personalized service can be introduced by combining mobile devices with context aware information during the process of it and feedback can be obtained after the end of the trip. Zhu et. al. modeled user in both geographical space and semantic space, defined Activity Pattern and extracted routes which matched individual's activity patterns from high similar users' trajectories to recommend top-k routes to a user [11]. Based on the Web GIS technology, liu et. al. designed a novel personalized smart system which was highlighted at spending the least travel cost to reach as many destinations as possible within a specified time period [12]. Hasuike et. al. introduced a Time-Expanded Network (TEN) to solve the problem of randomly changing of traveling and sightseeing times and selected the next sightseeing site through conditional probabilities calculated by current conditions, statistical and Web data [13]. Shen, Tong and Chen developed a two-step greedy-based heuristic algorithm to conduct strategic multiple-event planning for every user and consider the constraints of spatio-temporal conflicts and travel expenditure to address the data sparsity problem in destination prediction [14]. Xue et. al. proposed sub-trajectory synthesis method which decomposed historical trajectories into sub-trajectories comprising two adjacent locations and connected them into "synthesis" trajectories [15,16]. Su et. al. presented the CrowPlanner system which requested human workers to evaluate candidates routes recommended by different sources and methods and determined the best route based on the feedback of these workers [17]. Devasanthiya et. al. described a recommender system, which obtained textual messages, classified them using Rocchio classification and yielded the recommendation results using ontological specifications, to help travel agents in recommending tourism options to the customers [18]. The authors combined ontology-based semantic similarity measure to evaluate semantic similarity between items or to recommend personalized information [19,20].

In [21], the authors propose methods to provide passengers with useful information, the probability of taking a taxi and the average waiting time to facilitate their taxi-taking. In [22], Shen, Deng and Gao designed personalized attraction similarity model to suggest attractions by leveraging explicit user interaction and heterogeneous multi-modality travel information and refined the recommendation by adopting to context information such as the user location at a particular moment. Zheng explored the feasibility of promoting circuitous tourism through the recommendation of highly acclaimed tour routes [23].

2.2 LDA-based Recommendation

LDA model is a potential semantic analysis model in the field of text mining. It has been extended and applied to the field of recommendation. Chen et. al. used LDA model to divide the community to complete the community recommendation [24]. Liu et. al. modeled the user preference using LDA to improve the accuracy of CF recommendation [2]. Wang et. al. combined the traditional collaborative filtering algorithm with probabilistic topic model to recommend academic papers [3]. A questions recommendation method based on LDA topic model, which expressed the interests distributions through using LDA and calculated questions recommendation lists based on it, was proposed [25]. Li et. al. proposed a matrix generation model for cross domain collaborative filtering to fill the vacancy value for users and to model the user preference drift over time to get better recommendation [26]. In [27], based on an evaluation of similarity between the plot of a watched video by a user and a large amount of plots stored in a movie database, a plot-based recommendation system was proposed, which implemented and compared the two Topic Models, Latent Semantic Allocation (LSA) and LDA.

2.3 Combining Session-based Temporal Dynamic CF with LDA-based Recommendation

Session-based temporal Collaborative filtering model a user profile by dividing the user interaction history into stages. Yu and Zhu introduced an enhanced session-based temporal graph model considering three features to capture personal and temporal user interest and subsequently recommended personalized hashtags combining long-term and short-term user interest [4]. In [28], the authors presented incremental session-based collaborative filtering with forgetting mechanism in music recommendation systems, which considered music listened continuously and maintains the recent session. Ricardo et. al. defined the temporal information and the diversity of sessions and complete music recommendation using session-based collaborative filtering [29]. Xiang constructed a temporal graph to simultaneously model the user long-term and short-term preference [5]. Zheleva et. al. used LDA to build hierarchical graph after dividing a user interaction history into sessions [6]. Li et. al. divided the user interaction history into stages with a fixed time window and recommend the news group using the user long-term preference and specific news using the short-term preference [7]. Hong et. al. categorized the items to establish the user long-term preference, identify the user's current stage and provide the recommended list [8].

3. MODELING THE TOURIST PREFERENCE DRIFT OVER TIME

In this section, we firstly conduct studies on the characteristics analysis of tourism data in section A. In section B, based on the identified characteristics, we dynamically divide a tourist interaction history into sessions which alleviate the high sparsity of tourism data.

3.1 Characteristics of Tourism Data

Compared to other data sets, there are many characteristics of tourism data, such as the higher sparsity, temporal features of tourism, statistics characteristics on the age of tourists and the price of routes.

3.2 High Sparsity

The tourism data is more sparse compared to other standard data sets because the number of travel is very limited and the number of shopping or watching movies is very common. We use the percentage of tourists who travel times to show its sparsity. It is defined as (1)

$$P = \frac{N_i^{num}}{T_{num}}....$$
$$P = \frac{N_i^{num}}{T_{num}} \quad (1)$$

Where N_i^{num} represents the number of tourists who travel N_i times and T_{num} represents the total number of tourists. With the increase of N_i, the more sparse the data is ,the smaller the percentage is and most of tourists should be centered on the area of smaller N_i.

3.3 Temporal Features of Tourism

Tourism is an important way of leisure and entertainment, and it is easy to be influenced by the factors of the season and the leisure time. Assume that the tourist u_i has the entertainment time e_i and a year is divided into stages s_i $(1 \leq i \leq 12)$ by months. The probability of a tourism routes R_i in the stage s_i is selected by u_i and u_i may be to travel in s_i is defined respectively by (2) and (3)

$$p(R_i|u_i) \propto corr(e_i|s_i) \quad (2)$$
$$p(s_i|u_i) = a \quad (3)$$

where $corr(e_i|s_i)$ represents the correlations between the leisure and entertainment time e_i and the stage $s_i$the route R_i belongs to. a is a number that is either close to 0 or close to 1. It

represents that tourists are more willing to travel in their leisure and entertainment time and the time is relatively fixed in each year.

3.4 Statistics on the Age of Tourists and the Price of Routes

The age distribution of tourists is related to whether the tourists have spare time or have a strong economic capacity. The distribution of each age session when we divide the age of tourists into 6 sessions is defined by (4)

$$\mathrm{d}\left(\mathrm{age}_j\right) \propto \mathrm{e}_{\mathrm{u}_i} ? \mathrm{c}_{\mathrm{u}_i} (1 \leq \mathrm{j} \leq 6) \tag{4}$$

where $\mathrm{e}_{\mathrm{u}_i}$ and $\mathrm{c}_{\mathrm{u}_i}$ respectively represent the entertainment time and economic capacity of tourist u_i, age_j is the age session the tourist u_i belong to.

Unlike movies or shopping, the price is independent of the time, while the price of routes is often associated with the long of tourism time. The longer the tourism time is, the higher the price is. We split the price of all routes into 7 sections and then the influence of each price section is defined by (5)

$$\mathrm{f}\left(\mathrm{price}_k\right) \propto \frac{\mathrm{num}_{\mathrm{R}_k}}{\mathrm{num}_{\mathrm{Total}}} (1 \leq \mathrm{k} \leq 6) \tag{5}$$

Where $\mathrm{num}_{\mathrm{R}_k}$ is the number of routes whose price are in the section. $\mathrm{num}_{\mathrm{Total}}$ is the total number of all routes.

Therefore, we should take into account the characteristics of tourism data when the user preference is modeled to design a suitable recommendation algorithm and get better recommendation effect.

3.5 Modeling the Tourist Preference Drift over Time

1. *Division of Tourist Interaction History*

Session-based CF divide the user interaction history with fixed time window into sessions and the user profile is expressed as a sequence of these stages. However, the division with fixed time window is not suitable for tourism data because of the high sparsity of tourism data and relative fixed time slots of the tourists travel. That will lead to completely no tourism behavior in some stage if we divide it with fixed time window. So we should dynamic divide the tourist interaction history into stages with variable time window based on the characteristics of the actual interaction of each tourist.

Generally, we define a set of tourists $U = \{u_1, u_2, ..., u_m\}$, a set of routes $L = \{l_1, l_2, ..., l_n\}$, the interaction history of tourist u_i is H_{u_i}. First, we set the smallest temporal domain with the size of δ based on the average time interval of tourist u_i. Next, we select any record as a center point and compute the distance between it and other existing records. if the distance is greater than δ , the other record will be considered as a new center, otherwise, the two records will be merged into a clustering and calculate their center as a new center point. The distances between the new center point and the rest of records are recalculated and the process is continue until the results of any two adjacent point are not changed. Based on this process, H_{u_i}, the interaction history of u_i, is divided into stages $H_{u_i} = \{H_{u_i}^1, H_{u_i}^2, ..., H_{u_i}^{|H_{u_i}|}\}$, where $|H_{u_i}|$ is the number of center points, $H_{u_i}^t 1 \leq t \leq |H_{u_i}|$ is the records of the tourist u_i belong to the t^{th} clustering, $H_{u_i}^{|H_{u_i}|}$ is the latest one. Therefore, the interaction history of u_i will be dynamic divided into stages according to his own results that the different tourism records will be merged. We regard that the strategy our proposed not only can be applied to our data set, but also to other tourism data sets, and even to other types of data sets as long as the interaction history is segmented and data is sparse.

1. *Generation of Probabilistic Topic Distribution Based on LDA*

LDA as the language model is employed in recommendation. The general framework has a natural interpretation when dealing with users' preference data: the set of users define the corpus, each user is considered as a document, the items purchased are considered as words, the ratings are considered as the appeared frequency. In this paper, tourist records $H_{u_i}^t$ of the tourist u_i have the corresponding detailed tourism document $L_{u_i}^t$, so the LDA regard every document as a distribution of a group of topics to detect the probabilistic topic distribution for each tourist, and each topic is considered as the distribution of words about the description of some route. Therefore, the tourism document is firstly preprocessed by removing disable words and the low frequency words, and intelligently segmented based on forward iteration fine-grained segmentation. Then we can obtain the Polynomial distributions ϕ_{jk} and θ_{kl} of topic-word and topic- document using LDA which are represent by (6) and (7), respectively.

$$\phi_{jk} = \frac{n_{jk} + \beta_j}{\sum_{j=1}^{|L_j|} \left(n_{jk} + \beta_j\right)} \tag{6}$$

$$\theta_{kl} = \frac{m_{kl} + \alpha_k}{\sum_{k=1}^{K} \left(m_{kl} + \alpha_k\right)} \tag{7}$$

Where $|L_j|$ is the number of words in a document L_j, n_{jk} is the number of times that a word w_j is given to T_k, T_k is the k^{th} component of the topic vector $T = \{T_1, T_2, ..., T_K\}$, K is the number of topics, m_{kl} is the number of times that a document L_l is given to T_k, α and β are the super parameter of θ and ϕ, respectively. In practice, the default values of α and β are often set to $50/K$ and 0.01. So we can obtain the probabilistic topic distribution of the tourist u_i in t^{th} stage $P_{u_i}^t = (p_{u_i,1}^t, p_{u_i,2}^t, \cdots, p_{u_i,K}^t)^T$. The probabilistic topic distribution of the whole tourism history of the tourist u_i is expressed as $P_{u_i} = \{P_{u_i}^1, P_{u_i}^2, \cdots, P_{u_i}^{|H_{u_i}|}\}$.

(3) Tourist Preference Drift over Time

The probabilistic topic distribution of the later stage is more important than others because the later preference is more consistent with the current interest. So the probabilistic topic distributions of different stages have different weights to predict the tourist current interest. The preference drift over time is defined

as (8)

$$P_{u_i}^{|H_{u_i}|+1} = \sum_{t=1}^{|H_{u_i}|} \lambda_{u_i}^t P_{u_i}^t \tag{8}$$

Where $\lambda_{u_i}^t$ is the time decreasing weight of the tourist u_i in t^{th} stage. It is defined by (9)

$$\lambda_{u_i}^t = \frac{2t}{|H_{u_i}|\,(|H_{u_i}| - 1)} \tag{9}$$

It can be seen from (9) that $\lambda_{u_i}^t \in [0, 1]$ and the larger t represents the later stage of history. The higher value of $\lambda_{u_i}^t$ is, the greater contribution of probabilistic topic distribution is when modeling the tourist preference drift over time.

4. TOURISM ROUTES RECOMMENDATION

In this section, we firstly analyze the difficulty of neighbor selection in tourism data and propose a method of neighbor selection based on the tourist feature vector. Then, we can get the candidate set of routes based on the temporal features and recommend routes to the active tourist.

4.1 Selection of Tourist Neighbor

Due to the high sparsity of tourist data, the common routes between tourists are very few. Fig. 1 shows the change of the percentage of tourists who have the common routes. From Fig. 1 we can see that the common routes of over 95% tourists are less than 3, and the proportion of its within one month is slightly higher. That shows that the number of common routes of all tourists are very few, less than 5 times, which lead to the difficulty of selection of nearest neighbors whom the time of the common routes are relatively close. So we can use these characteristics of tourism data in the course of recommending routes.

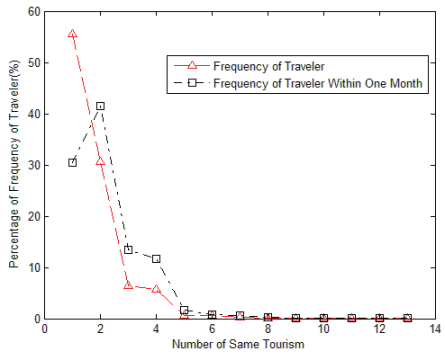

Figure 1 Percentage of Tourists with Common Routes.

Based on a detailed analysis of the characteristics of the tourism data and the segmentation of these attributes in section III, the travel time of tourists is divided into four periods based on spring, summer, autumn and winter. The age of them is divided into six stages. We divide the price of routes into seven sections. We establish the d dimensional feature vector $V_{u_i} = \{v_{u_i}^1, v_{u_i}^2, \cdots, v_{u_i}^d\}$ for tourist u_i, each one can be expressed as a discrete value.

$$v_{u_i}^q = \begin{cases} 1 \\ 0 \end{cases}$$

Then the similarities between tourists are calculated based on the feature vectors of tourists and is defined as (10)

$$sim(u_i, u_s) = \sqrt{\sum_{q=1}^{d} (v_{u_i}^q - v_{u_s}^q)^2} \tag{10}$$

4.2 Generation of Candidate Routes

We can see that the tourists are more willing to travel in the relative specific month in each year based on the analysis of section III and the time of the common routes the is relatively close based on IV. Based on these analysis, we set the three tuple of tourist $\left\langle u_i, m_f, L_{u_i}^{m_f} \right\rangle 1 \le f \le 12$ after a year is divided into 12 stages in accordance with months, which represents the routes set $L_{u_i}^{m_f}$ of the tourist u_i in the month of m_f. The sets of temporal neighbors and candidate routes are obtained from (11) and (12), respectively.

$$N_{u_i}^{tem} = \{u_{ts} \left| L_{u_i}^{m_f} \cap L_{u_{ts}}^{m_f} \neq \Phi \,, u_{ts} \in N_{u_i}\} \tag{11}$$

$$S_{u_i} = \{L_c \,| L_c \in L_{u_a} ? L_c \notin L_{u_i}, u_a \in N_{u_i}^{tem}\} \tag{12}$$

Where N_{u_i} is the neighbors obtained from (10).

4.3 Tourism Routes Recommendation

The probabilistic topic distribution P_{L_l} for each route in the candidate set S_{u_i} of tourist u_i is obtained using LDA, and the preference drift over time of tourist u_i in the stage $|H_{u_i}| + 1$ is calculated by using (8), which are K dimensional vectors in the probabilistic topic space. The common similarity measure used in recommender systems are cosine similarity, adjusted cosine similarity and pearson correlation-based similarity. Cosine similarity can measure the similarity between vectors using the cosine value of the angle of them. So we compute the similarity between the probabilistic topic distribution P_{L_l} and $P_{u_i}^{|H_{u_i}|+1}$ using Cosine similarity to obtain the correlation degree between the user predicted preference and the route by (13).

$$sim(P_{u_i}^{|H_{u_i}|+1}, P_{L_l}) = \frac{P_{u_i}^{|H_{u_i}|+1} \bullet P_{L_l}}{||P_{u_i}^{|H_{u_i}|+1}||\,||P_{L_l}||} \tag{13}$$

Where $L_l \in S_{u_i}$, we rank the similarities and recommend top-k routes to the active tourist.

5. EXPERIMENTS AND RESULTS

In this section, we conduct comprehensive experimental evaluation to show efficacy and effectiveness of our proposed method.

First, we introduce the tourism dataset used in our paper and metric for performance comparison which is defined by us based on the tourism characteristics. Then, we present the experimental results including key parameters finding, model validation and comparison studies.

5.1 Dataset

The real tourism data comes from a tourism company belong to Xiamen airlines. The data set contains 732019 travel records from January 2009 to October 2014. In this paper, we extract 25717 records from 4737 tourists on 1436 routes. The tourists with less than 3 times records are removed. Currently, the data set is not published on the Internet because of the privacy of tourists.

5.2 Metric

Precision, Recall and F-score are metrics to evaluate the performance of top-k recommendation. In the experimental process, we treat the first $\left|H_{u_i}\right| - 1$ times data for each tourist as the training set, the last $\left|H_{u_i}\right|$ data as test set. Because of the high sparsity, the precision is almost 0 or $1/k$ and the recall is 0 or 1 to evaluate the recommendation results because the number of recommend correctly route is nearly 0 or 1. In this paper, we propose precision coverage as the evaluation metric, which is defined as (14)

$$P_{cov} = \frac{\sum_{u_i \in U} \rho_{u_i}}{|U|} \tag{14}$$

Where $|U|$ is the number of tourists, and ρ_{u_i} is defined by (15)

$$\rho_{u_i} = \begin{cases} 1 & L_{u_i}^{|H_{u_i}|} \in S_{u_i}^{Top-k} \\ 0 & L_{u_i}^{|H_{u_i}|} \notin S_{u_i}^{Top-k} \end{cases} \tag{15}$$

$L_{u_i}^{|H_{u_i}|}$ is the actual route of tourist u_i in stage $\left|H_{u_i}\right|$, $S_{u_i}^{Top-k}$ is the set of top-k routes recommended to the active tourist u_i.

5.3 Experimental Results and Analysis

1. *Analysis of Tourism Data*

We compare the tourism data set to a standard movie recommendation data set (Movelens). In order to make a better comparison, the number of ratings is 10 times the number of travel times. The result is shown in Fig. 2. From Fig. 2 it can be seen that the percentage of tourists has declined rapidly with the increase of the times of tourism, and more than 95% tourists are less than 10 times. Its percentage also decreased with the increase of the number of watched movies on Movielens, but the speed is significantly slower than the tourism data and the gap is more obvious with the larger number of them.

We statistic on the tourism months of all tourists in Fig. 3. Figure 3 shows that tourists are more willing to travel in the spring and autumn when have a pleasant climate. The tourism time distribution is more concentrated for each tourist. Fig. 4 shows the month distribution of tourists. From Fig. 4 we can see that the tourism time of more than 70% of the tourists is concentrated in 4 months, that shows that the tourism time for each tourist is a relatively fixed time in each year. So the time factor should be considered when we recommend routes to the users.

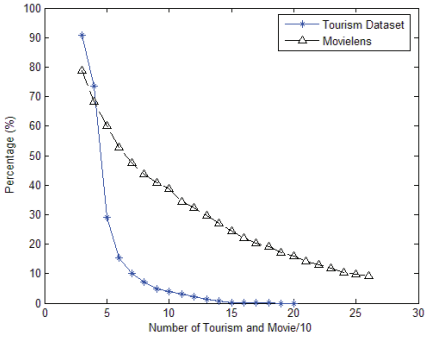

Figure 2 Comparison with Sparsity.

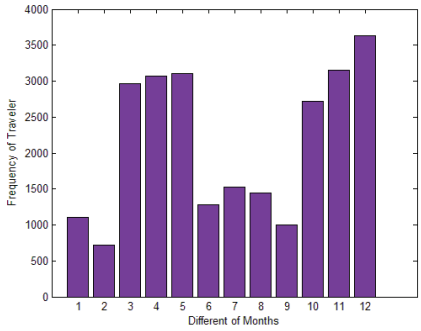

Figure 3 Statistics on Tourism Month.

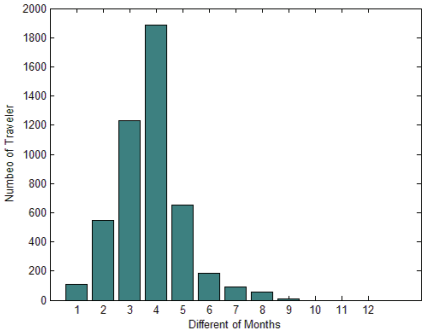

Figure 4 Tourism Months Distribution of Tourists.

Figure 5 shows that the age of the main force of tourists is distributed in the 1-18, 26-35 and 36-50 years old, which exceed 70% of the total tourists. It is most likely because that students whose age are 1-18 years old have a lot of spare time, such as winter and summer vacation, to follow their parents or their partners to travel. The groups whose age in 26-35 and 36-50 years old have a strong economic capacity, and tourism has become a way of their leisure and entertainment.

The price of selected by tourists is shown in Fig. 6. From Fig. 6 it can be seen that the number of tourists is reduce rapidly when the price of routes is increase. About 70% of the tourists choose the routes where the price is below 500. The proportion of tourists who select the price of routes between 500-2000 is

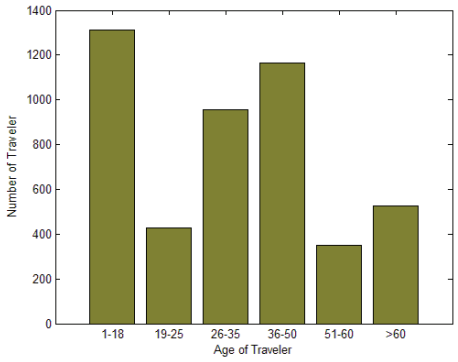

Figure 5 Distribution of the Ages of Tourists.

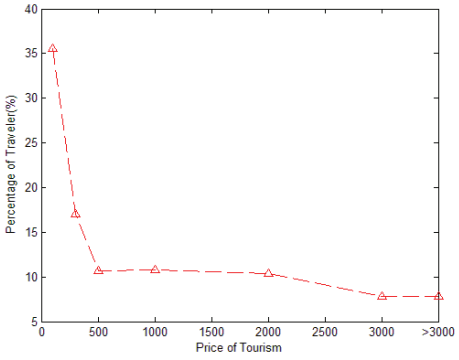

Figure 6 Influence of Price of Routes.

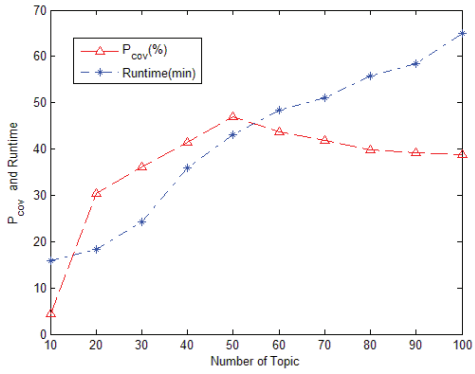

Figure 7 Tuning the Number of Topics.

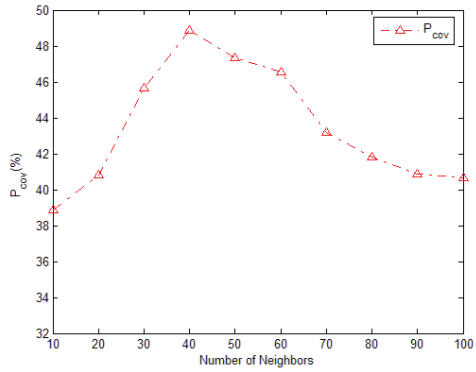

Figure 8 Influence of the Number of Neighbors.

flat, while the percentage of tourists selected 3000 and more than 3000 is lower than others and they are very close. Therefore, we can see that people prefer the routes of which the price is cheap and the time is short. The influence of price becomes smaller when the price is reach to a certain value.

(2)Selection of the Number of Topic

The optimal number of topics K used in the LDA-based recommendation is often not learned from the data, but is predefined because the topic is latent variable. We use precision coverage as the criteria with different K values, meanwhile, record the running time because Accuracy and computational complexity are the two criteria when we decided the number of topics. The complexity of recommendation is higher because of the more topics, which lead to the larger computation. The results are shown in Fig. 7. We can see that the precision coverage is increase and then decrease with the increase of topics, while the running time is always increasing. The threshold values that lead to better precision-coverage results and smaller computational complexity is 50, which is the optimal number of topics we selected.

(3) Selection of the Number of Neighbors n

In all the neighborhood-based methods, the number of similar neighbors n is very important. We calculate precision coverage for the active user with different value of n, as shown in Fig. 8.

We notice that Precision coverage increase and then decreases when the number of neighbors is increasing. That is because the common routes are always less if the neighbors are few, which lead to a smaller candidate routes set, while the similarities between the active user and neighbors become poor if the number of it is too large, which lead to the larger difference between the candidate routes and the actual route of active tourist. So in the following experiments, we will use 50 as the optimal number of neighbors for our approach.

5.4 Comparison with Other Methods

To evaluate the effectiveness of our method, we compare our method (called TLDA) with several representative baselines: 1)UCF (user-based collaborative filtering) [30]: a representative of user-based collaborative filtering; 2) LDA [1]: a method that the user preference is modeled using LDA and recommend items to the active user based on the user profile; 3) ItemRank [31]: a method that the association graph of routes is established to rank them using random walk. In the experiment, the LDA parameters are $\alpha = 50/K$ and $\beta = 0.01$, the restart probability is 0.15, and the other parameters take the optimal values for each method. The comparison results are shown in Fig. 9.

We observe that (1) TLDA, LDA and ItemRank methods outperform UCF. That is because LDA-based methods (TLDA and LDA) are think that the users' interest influenced by potential factors which deeply reflect and model interaction relationships between users and items, which can better capture the user preference, and ItemRank increases the recommendation diversity using random walk to prevent over filtering, but UCF only reply on rating information, which is less informative. (2) LDA and Itemrank perform similarity, and The performance of LDA is slightly superior to that of ItemRank, which is straightforward since LDA characterize the user preference using latent factors

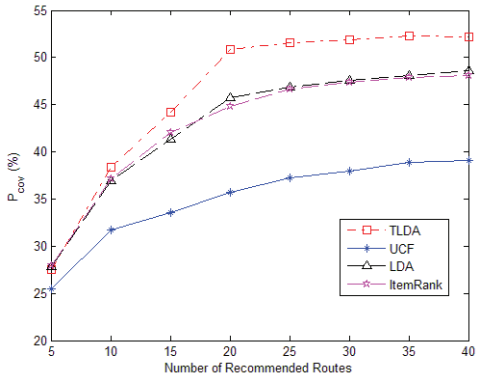

Figure 9 Comparison with Other Methods.

and recommend items to the active user based on it, which better predict the trend of profile of user, while the ItemRank only use the associations between items. (3) The performance of TLDA is better than that of all other methods. The reason is that TLDA not only models the user preference drift over time using LDA, but also takes temporal context, tourist age, route price and travel season into account when we select neighbors for the active user.

6. CONCLUSIONS

In this paper, we propose a novel method that automatically divides a tourist's interaction history into sessions with variable sizes based on its actual travel data. LDA is then used to model tourists' preference in different sessions and the time decreasing weights are used to measure their importance to capture their dynamic preference, which is easy to mining the preference changes for each tourist on topic and predict the tourist's preference trend on routes. We take the temporal context, tourist age, route price and travel season into account when the neighbors are selected to complete recommendation. The method we proposed not only mining and adapt to the high sparsity, temporal, seasonal and price characteristics, but also alleviate the high sparsity because a tourist interaction is divided into sessions. The experimental results on real tourism data set show that our method not only can dig out the tourist preference drift over time and but also achieve better recommendation performance.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (61301245; U1533104), Hebei Province Natural Science Foundation (No. E2016202341), Tianjin Natural Science Foundation (14JCZDJC32500) and the Fundamental Research Funds for the Central Universities (ZXH2012P009)

REFERENCES

1. M. Morzy. "Cluster-based analysis and recommendation of sellers in online auctions". Computer Systems Science and Engineering, 2007,Vol 22,Issue 5,pp:279-287.
2. Q Liu, E Chen, H Xiong,H Q Ding and J Chen. "Enhancing collaborative filtering by user interest expansion via personalized ranking". IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2012, Vol 42, Issue 1, pp:218-233.
3. C Wang and D M Blei. "Collaborative topic modeling for recommending scientific articles". Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011, pp: 448-456.
4. J Yu, T Zhu. "Combining long-term and short-term user interest for personalized hashtag recommendation". Frontiers of Computer Science, 2015,Vol 9, Issue 4, pp: 608-622.
5. L Xiang, Q Yuan, S Zhao, L Chen, X Zhang, et al. "Temporal recommendation on graphs via long-and short-term preference fusion". Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010, pp:723-732.
6. E.Zheleva, J.Guiver, R.E.Mendes and N.Milic-Frayling."Statistical models of music-listening sessions in social media." Proceeding of 19th International Conference on World Wide Web, 2010, pp:101-1028.
7. L Li, L Zheng, F Yang and T Li. "Modeling and broadening temporal user interest in personalized news recommendation". Expert Systems with Applications, 2,14, Vol 41, Issue 7, pp:3168-3177.
8. Wenxing Hong, Lei Li and Tao Li. "Product recommendation with temporal dynamics". Expert Systems with Applications,2012, Vol 39, Issue 16, pp:12398-12406.
9. Z Q He, Z Y Wu, B C Zhou, L Xu and W F Zhang. "Tourist routs recommendation based on latent dirichlet allocation model". Proceedings of 2015 12th Web Information System and Application Conference, 2015, pp: 201-206.
10. X Chen and L Q Zhou. "Design and implementation of an intelligent system for tourist routes recommendation based on Hadoop". Proceedings of 2015 IEEE 6th International Conference on Software Engineering and Service Sciences, 2015 November, pp:774-778.
11. L C Zhu, Z J Li and S X Jiang. "LBSN-based personalized routes recommendation". Applied Mechanics and Materials, 2014, Vol 644-650, pp:3230-3234.
12. H L Liu, J H Li and J Peng. "A novel recommendation system for the personalized smart tourism route: Design and implementation". Proceedings of 2015 IEEE 14th International Conference on Cognitive Informatics and Cognitive Computing, 2015, pp: 291-296.
13. T Hasuike, H Katagiri, H Tsubaki and H Tsuda. "A Route Recommendation System for Sightseeing with Network Optimization and Conditional Probability". Proceedings of 2015 IEEE International Conference on Systems, Man, and Cybernetics, 2015, pp: 2672-2677.
14. J She, Y Tong and L Chen. "Utility-aware event-participant planning". Proceedings of the 36th ACM International Conference on Management of Data , 2015 May, PP:1629-1643.
15. A Y Xue, R Zhang, Y Zheng, X Xie, J Huang and Z Xu. "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction". IEEE International Conference on Data Engineering , 2013, pp;254-265.
16. A Y Xue, J Qi, X Xie, R Zhang, J Huang, and Y Li. "Solving the data sparsity problem in destination prediction". VLDB Journal, 2015, Vol 24 , Issue 2, pp: 219-243.
17. H Su, K Zheng, J Huan, T Liu, H Wang and X Zhou. "A crowd-based route recommendation system-CrowdPlanner". Proceedings of 2014 IEEE 30th International Conference on Data Engineering, 2014, pp: 1178-1181.
18. C Devasanthiya, S Vigneshwari, and J Vivek. "An enhanced tourism recommendation system with relevancy feedback mechanism and ontological specifications". Advances in Intelligent Sys-

tems and Computing, 2016,Vol 398, pp: 281-289.
19. M Al-Hassan, H Lu and J Lu. “A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system”. Decision Support Systems, 2015,Vol 72, pp: 97-109.
20. M. Sohn, H. Kim and H.J. Lee. “Personalized recommendation framework based on CBR and CSP using ontology in a ubiquitous computing environment”. Computer Systems Science and Engineering, 2012,Vol 27,Issue 6,pp:415-430.
21. Y.F. Chen, X. Zhao, J.Y.Tang, W.M. Zhang and H.C. Shang. “Taxi-taking recommendation using real-time trajextories: an online query based approach”. Computer Systems Science and Engineering, 2016,Vol 31,Issue 2,pp:117-125.
22. J Shen, C Deng and X Gao. “Attraction recommendation: Towards personalized tourism via collective intelligence”. Neurocomputing, 2016, Vol 173, pp: 789-798.
23. M C Zheng. “How a map with a tour route recommendation promotes circuitous tourism”. Journal of Asian Architecture and Building Engineering, 2015,Vol 14, Issue 2, pp: 363-370.
24. W Y Chen, J C Chu, J Luan, H Bai, Y Wang and E Y Chang. “Collaborative filtering for orkut communities: discovery of user latent behavior”. Proceedings of the 18th International Conference on World Wide Web. ACM, 2009, pp:681-690.
25. C Wang, L Cu, B Yang and X Wu. “Question recommendation mechanism under Q&A Community based on LDA Model”. Open Cybernetics and Systemics Journal, 2014, Vol 8, pp: 645-650.
26. B Li, X Q Zhu, R J Li, C Q Zhang, X Y Xue and X D Wu. “Cross-Domain Collaborative Filtering over Time”. Proceeding of the 22th International Joint Conference on Artificial Intelligence, 2012, PP:2293-2298.
27. S Bergamaschi, L Po. “Comparing LDA and LSA topic models for content-based movie recommendation systems”. Lecture Notes in Business Information Processing, 2015, Vol 226, pp: 247-263.
28. U Suksawatchon,S Darapisut, J Suksawatchon. “Incremental session based collaborative filtering with forgetting mechanisms”. Proceeding of 2015 19th International Computer Science and Engineering Conference: Hybrid Cloud Computing: a New Approach for Big Data Era, February 8, 2016.
29. D.Ricardo and M. J. Fonseca. “Improving music recommendation in session-based collaborative filtering by using temporal context”. Proceeding. of IEEE 25th International on Tools with Artificial Intelligence, 2013, pp:783-788.
30. P Resnick, N Iacovou, M Suchak, P Bergstron and J Riedl. “GroupLens: an open architecture for collaborative filtering of netnews”. Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work. ACM, 1994, pp:175-186.
31. M Gori and A Pucci. ItemRank: “a random-walk based scoring algorithm for recommender engines”. Proceedings of the 20th International Joint Conference on Artifical Intelligence. Morgan Kaufmann Publishers Inc.2007, pp:2766-2771.