A Cocktail Approach for Travel Package Recommendation

Qi Liu, Enhong Chen, *Senior Member*, *IEEE*, Hui Xiong, *Senior Member*, *IEEE*, Yong Ge, Zhongmou Li, and Xiang Wu

Abstract—Recent years have witnessed an increased interest in recommender systems. Despite significant progress in this field, there still remain numerous avenues to explore. Indeed, this paper provides a study of exploiting online travel information for personalized travel package recommendation. A critical challenge along this line is to address the unique characteristics of travel data, which distinguish travel packages from traditional items for recommendation. To that end, in this paper, we first analyze the characteristics of the existing travel packages and develop a tourist-area-season topic (TAST) model. This TAST model can represent travel packages and tourists by different topic distributions, where the topic extraction is conditioned on both the tourists and the intrinsic features (i.e., locations, travel seasons) of the landscapes. Then, based on this topic model representation, we propose a cocktail approach to generate the lists for personalized travel package recommendation. Furthermore, we extend the TAST model to the tourist-relation-area-season topic (TRAST) model for capturing the latent relationships among the tourists in each travel group. Finally, we evaluate the TAST model, the TRAST model, and the cocktail recommendation approach on the real-world travel package data. Experimental results show that the TAST model can effectively capture the unique characteristics of the travel data and the cocktail approach is, thus, much more effective than traditional recommendation techniques for travel package recommendation. Also, by considering tourist relationships, the TRAST model can be used as an effective assessment for travel group formation.

Index Terms—Travel package, recommender systems, cocktail, topic modeling, collaborative filtering

1 Introduction

As an emerging trend, more and more travel companies provide online services. However, the rapid growth of online travel information imposes an increasing challenge for tourists who have to choose from a large number of available travel packages for satisfying their personalized needs. Moreover, to increase the profit, the travel companies have to understand the preferences from different tourists and serve more attractive packages. Therefore, the demand for intelligent travel services is expected to increase dramatically.

Since recommender systems have been successfully applied to enhance the quality of service in a number of fields [2], [15], [37], it is natural choice to provide travel package recommendations. Actually, recommendations for tourists have been studied before [1], [4], [8], and to the best of our knowledge, the first operative tourism recommender system was introduced by Delgado and Davidson [11]. Despite of the increasing interests in this field, the problem of leveraging unique features to distinguish personalized travel package recommendations from traditional recommender systems remains pretty open.

Indeed, there are many technical and domain challenges inherent in designing and implementing an effective recommender system for personalized travel package recommendation. First, travel data are much fewer and sparser than traditional items, such as movies for recommendation, because the costs for a travel are much more expensive than for watching a movie [14], [43]. Second, every travel package consists of many landscapes (places of interest and attractions), and, thus, has intrinsic complex spatio-temporal relationships. For example, a travel package only includes the landscapes which are geographically colocated together. Also, different travel packages are usually developed for different travel seasons. Therefore, the landscapes in a travel package usually have spatial-temporal autocorrelations. Third, traditional recommender systems usually rely on user explicit ratings. However, for travel data, the user ratings are usually not conveniently available. Finally, the traditional items for recommendation usually have a long period of stable value, while the values of travel packages can easily depreciate over time and a package usually only lasts for a certain period of time. The travel companies need to actively create new tour packages to replace the old ones based on the interests of the tourists.

To address these challenges, in our preliminary work [25], we proposed a cocktail approach on personalized travel package recommendation. Specifically, we first analyze the key characteristics of the existing travel packages. Along this line, travel time and travel destinations are divided into different seasons and areas. Then, we

- *Q. Liu, E. Chen, and X. Wu are with the School of Computer Science and Technology, University of Science and Technology of China, 502 DianSan Building, West Campus USTC, HuangShan Road, Hefei, Anhui 230027, China. E-mail: liuqiah@gmail.com, cheneh@ustc.edu.cn, wux@mail.ustc.edu.cn.*
- *H. Xiong and Z. Li are with the Management Science and Information Systems Department, Rutgers Business School, Rutgers, the State University of New Jersey, 1 Washington Park, Newark, NJ 07102. E-mail: hxiong@rutgers.edu, mosesli@pegasus.rutgers.edu.*
- *Y. Ge is with the Department of Computer Science, University of North Carolina, 9201 University City Blvd, Charlotte, NC 28223-0001. E-mail: yong.ge@uncc.edu.*

Manuscript received 13 Jan. 2012; revised 14 Aug. 2012; accepted 16 Nov. 2012; published online 28 Nov. 2012.
Recommended for acceptance by T. Sellis.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2012-01-0028.
Digital Object Identifier no. 10.1109/TKDE.2012.233.

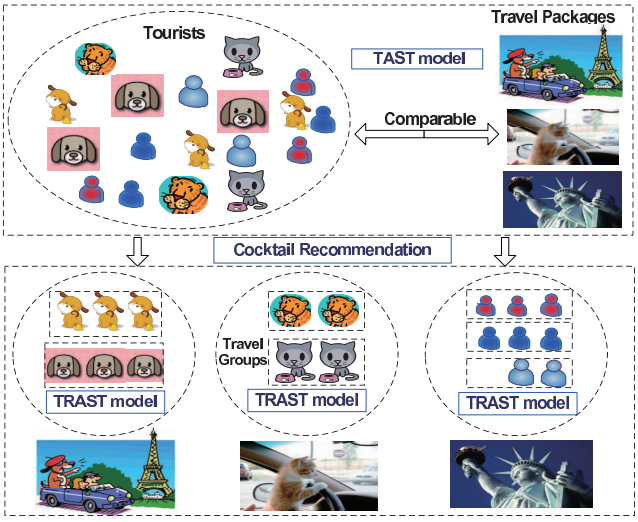


Fig. 1. An illustration of the paper contribution.

develop a tourist-area-season topic (TAST) model, which can represent travel packages and tourists by different topic distributions. In the TAST model, the extraction of topics is conditioned on both the tourists and the intrinsic features (i.e., locations, travel seasons) of the landscapes. As a result, the TAST model can well represent the content of the travel packages and the interests of the tourists. Based on this TAST model, a cocktail approach is developed for personalized travel package recommendation by considering some additional factors including the seasonal behaviors of tourists, the prices of travel packages, and the cold start problem of new packages. Finally, the experimental results on real-world travel data show that the TAST model can effectively capture the unique characteristics of travel data and the cocktail recommendation approach performs much better than traditional techniques.

In this paper, we further study some related topic models of the TAST model, and explain the corresponding travel package recommendation strategies based on them. Also, we propose the tourist-relation-area-season topic (TRAST) model, which helps understand the reasons why tourists form a travel group. This goes beyond personalized package recommendations and is helpful for capturing the latent relationships among the tourists in each travel group. In addition, we conduct systematic experiments on the real-world data. These experiments not only demonstrate that the TRAST model can be used as an assessment for travel group automatic formation but also provide more insights into the TAST model and the cocktail recommendation approach. In summary, the contributions of the TAST model, the cocktail approaches, and the TRAST model for travel package recommendations are shown in Fig. 1, where each dashed rectangular box in the dashed circle identifies a travel group and the tourists in the same travel group are represented by the same icons.

2 Concepts and Data Description

In this section, we first introduce the basic concepts, and then describe the recommendation scenario of this study. Finally, we provide the detailed information about the unique characteristics of travel package data.

Fig. 2. An example of the travel package, where the landscapes are represented by the words in red.

Definition 1. *A* travel package *is a general service package provided by a travel company for the individual or a group of tourists based on their travel preferences. A package usually consists of the landscapes and some related information, such as the price, the travel period, and the transportation means.*

Specifically, the travel topics *are the themes designed for this package, and the* landscapes *are the travel places of interest and attractions, which usually locate in nearby* areas.

Following Definition 1, an example document for a package named "*Niagara Falls Discovery*" from the STA Travel[1] is shown in Fig. 2. It includes the travel topics (tour style), travel days, price, travel area (the northeastern US), and landscapes (e.g., Niagara Falls), and so on. Note that different packages may include the same landscapes and each landscape can be used for multiple packages. Meanwhile, for some reasons, the tourists for each individual package are often divided into different travel *groups* (i.e., traveling together). In addition, each package has a travel schedule and most of the packages will be traveled only in a given time (*season*) of the year, i.e., they have strong seasonal patterns. For example, the "*Maple Leaf Adventures*" is usually meaningful in Fall.

In this paper, we aim to make personalized travel package recommendations for the tourists. Thus, the users are the tourists and the items are the existing packages, and we exploit a real-world travel data set provided by a travel company in China for building recommender systems. There are nearly 220,000 expense records (purchases of individual tourists) starting from January 2000 to October 2010. From this data set, we extracted 23,351 useful records of 7,749 travel groups for 5,211 tourists from 908 domestic and international packages in a way that each tourist has traveled at least two different packages. The extracted data contain 1,065 different landscapes located in 139 cities from 10 countries. On average, each package has 11 different landscapes, and each tourist has traveled 4.4 times.

As illustrated in our preliminary work [25], there are some unique characteristics of the travel data. First, it is very sparse, and each tourist has only a few travel records. The extreme sparseness of the data leads to difficulties for using traditional recommendation techniques, such as collaborative filtering. For example, it is hard to find the credible nearest neighbors for the tourists because there are very few cotraveling packages.

1. STA Travel, URL: http://www.statravel.com/.

TABLE 1
Mathematical Notations

Notation	Description
$U = \{U_1, U_2, ..., U_M\}$	the set of tourists
$S = \{S_1, S_2, ..., S_J\}$	the set of seasons
$P = \{P_1, P_2, ..., P_N\}$	the set of packages
$T = \{T_1, T_2, ..., T_Z\}$	the set of topics
$A = \{A_1, A_2, ..., A_O\}$	the set of different areas
$P^{'} = \{P_1^{'}, P_2^{'}, ..., P_{D^{'}}^{'}\}$	packages for travel logs
$P^{''} = \{P_1^{''}, P_2^{''}, ..., P_{D^{'}}^{''}\}$	packages for travel group logs
$L_{A_i} = \{L_{A_i 1}, ..., L_{A_{i\|A_i\|}}\}$	landscape set for area A_i
$L_{P_i^{'}} = \{L_{P_i^{'} 1}, ..., L_{P_{i\|P_i^{'}\|}^{'}}\}$	landscapes for the package $P_i^{'}$
$L_{P_i^{''}} = \{L_{P_i^{''} 1}, ..., L_{P_{i\|P_i^{''}\|}^{''}}\}$	landscapes for the package $P_i^{''}$

Second, the travel data has strong time dependence. The travel packages often have a life cycle along with the change to the business demand, i.e., they only last for a certain period. In contrast, most of the landscapes will still be active after the original package has been discarded. These landscapes can be used to form new packages together with some other landscapes. Thus, we can observe that the landscapes are more sustainable and important than the package itself.

Third, landscape has some intrinsic features like the geographic location and the right travel seasons. Only the landscapes with similar spatial-temporal features are suitable for the same packages, i.e., the landscapes in one package have spatial-temporal autocorrelations and follow the first law of geography-everything is related to everything else, but the nearby things are more related than distant things [10]. Therefore, when making recommendations, we should take the landscapes' spatial-temporal correlations into consideration so as to describe the tourists and the packages precisely.

Fourth, the tourists will consider both time and financial costs before they accept a package. This is quite different from the traditional recommendations where the cost of an item is usually not a concern. Thus, it is very important to profile the tourists based on their interests as well as the time and the money they can afford. Since the package with a higher price often tends to have more time and vice versa, in this paper we only take the price factor into consideration.

Fifth, people often travel with their friends, family, or colleagues. Even when two tourists in the same travel group are totally strangers, there must be some reasons for the travel company to put them together. For instance, they may be of the same age or have the same travel schedule. Hence, it is also very important to understand the relationships among the tourists in the same travel group. This understanding can help to form the travel group.

Last but not least, few tourist ratings are available for travel packages. However, we can see that every choice of a travel package indicates the strong interest of the tourist in the content provided in the package.

In summary, these characteristics bring in three major challenges. First, how to compare the interests of tourists and the content of the travel package; second, how to make package recommendations for each tourist; third, how to capture the tourist relationships to form a travel group. As a result, it is necessary to develop more suitable approaches for travel package recommendation.

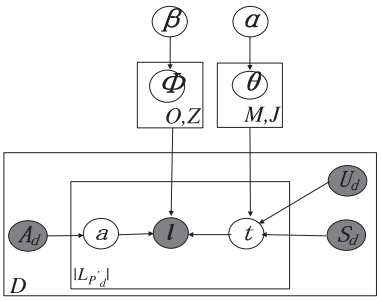


Fig. 3. TAST: A graphical model.

3 The TAST Model

In this section, we show how to represent the packages and tourists by a topic model, like the methods in [5], [29], and [36] based on Bayesian networks, so that the similarity between packages and tourists can be measured. Table 1 lists some mathematical notations in this paper.

3.1 Topic Model Representation

When designing a travel package, we assume that the people in travel companies often consider the following issues. First, it is necessary to determine the set of target tourists, the travel seasons, and the travel places. Second, one or multiple travel topics (e.g., *"The Sunshine Trip"*) will be chosen based on the category of target tourists and the scheduled travel seasons. Each package and landscape can be viewed as a mixture of a number of travel topics. Then, the landscapes will be determined according to the travel topics and the geographic locations. Finally, some additional information (e.g., price, transportation, and accommodations) should be included. According to these processes, we formalize package generation as a *W*hat-*W*ho-*W*hen-*W*here (4*W*) problem. Here, we omit the additional information and each W stands for the travel topics, the target tourists, the seasons, and the corresponding landscape located areas, respectively. These four factors are strongly correlated.

Formally, we reprocess the generation of a package in a topic model style, where we treat it mainly as a landscape drawing problem. These landscapes for the package are drawn from the landscape set one by one. For choosing a landscape, we first choose a topic from the distribution over topics specific to the given tourist and season, then the landscape is generated from the chosen topic and travel area. We call our model for package representation as the TAST model. Please note that, a topic mentioned in TAST is different from a real topic, where the former one is a latent factor extracted by topic model, while the latter one is an explicit travel theme identified in the real world, and latent topics are used to simulate real topics. Without loss of generality, we use *travel topic* and *topic* to stand for the real and latent topic, respectively.

Mathematically, the generative process corresponds to the hierarchical Bayesian model for TAST is shown in Fig. 3,

where shaded and unshaded variables indicate observed and latent variables, respectively. The TAST model follows the similar Dirichlet distribution assumptions as [5], [29], [36], and here landscapes are the *"tokens"* for topic modeling. In TAST model, the notation $P_d^{'}$ is different from P_d, where P_d is the ID for a package in the package set while $P_d^{'}$ stands for the package ID of one travel log, and each travel log can be distinguished by a vector of three attributes $\langle P_d^{'}, U_d, timestamp\rangle$, where the *timestamp* can be further projected to a season S_d and $P_d^{'} = \langle L_{P_d^{'}}, A_d, price^2\rangle$. Specifically, in Fig. 3, each package $P_d^{'}$ is represented as a vector of $|L_{P_d^{'}}|$ landscapes where landscape l is chosen from one area a and $a \in A_d$ (A_d includes the located area(s) for $P_d^{'}$) and (U_d, S_d) is the specific tourist-season pair. t is a topic which is chosen from the set T with Z topics. θ and ϕ correspond to the topic distribution and landscape distribution specific to each tourist-season pair and area-topic pair, respectively, where α and β are the corresponding hyperparameters.

The distributions, such as θ and ϕ, can be extracted after inferring this TAST model (*"invert"* the generative process and *"generate"* latent variables). The general idea is to find a latent variable (e.g., topic) setting so as to get a marginal distribution of the travel log set $P^{'}$:

$$p(P^{'}|\alpha,\beta,U,S,A) = \int\int \prod_{m=1}^{M}\prod_{j=1}^{J} p(\theta_{mj}|\alpha) \prod_{o=1}^{O}\prod_{k=1}^{Z} p(\phi_{ok}|\beta) \prod_{d=1}^{D}\prod_{i=1}^{|L_{P_d^{'}}|}\sum_{t_{di}=1}^{Z}(p(t_{di}|\theta_{U_d S_d}) \sum_{a_{di}\in A_d}(p(a_{di}|A_d)p(l_{di}|\phi_{a_{di}t_{di}})))d\phi d\theta.$$

3.2 Model Inference

While the inference on models in the LDA family cannot be solved with closed-form solutions, a variety of algorithms have been developed to estimate the parameters of these models. In this paper, we exploit the Gibbs sampling method [18], a form of Markov chain Monte Carlo, which is easy to implement and provides a relatively efficient way for extracting a set of topics from a large set of travel logs. During the Gibbs sampling, the generation of each landscape token for a given travel log depends on the topic distribution of the corresponding tourist-season pair and the landscape distribution of the area-topic pair. Finally, the posterior estimates of θ and ϕ given the training set can be calculated by

$$\hat{\theta_{mjt}} = \frac{\alpha_t + n_{mjt}}{\Sigma_{k=1}^{Z}(\alpha_k + n_{mjk})}, \quad \hat{\phi_{okl}} = \frac{\beta_l + m_{okl}}{\Sigma_{q=1}^{|A_o|}(\beta_q + m_{okq})}, \qquad (1)$$

where $|A_o|$ is the number of landscapes in area A_o, n_{mjt} is the number of landscape tokens assigned to topic T_t and tourist-season pair (U_m, S_j), and m_{okl} is the number of tokens of landscape L_l assigned to area-topic pair (A_o, T_k). Let us take the topic assignment for *"Central Park"* as an example, in each iteration, the topic assignment of one *"Central"* token depends on not only the topics of the landscapes traveled by the tourist in the given season but also the topics of the other landscapes located nearby. Meanwhile, many other posterior probabilities can also be estimated, for example, the topic distribution of tourist U_i and package P_i:

$$\vartheta_{ij}^{U} = \frac{\alpha_j + \Sigma_{s=1}^{J} n_{isj}}{\Sigma_{k=1}^{Z}(\alpha_k + \Sigma_{s=1}^{J} n_{isk})}, \quad \vartheta_{ij}^{P} = \frac{\alpha_j + h_{ij}}{\Sigma_{k=1}^{Z}(\alpha_k + h_{ik})}, \qquad (2)$$

where h_{ij} is the number of the landscape tokens in package P_i and these tokens are assigned to topic T_j.

After Gibbs sampling, all the tourists and packages are represented by the Z entry topic distribution vectors (Z, the number of topics, is usually in the range of [20, 100]). For example, a tourist, who traveled *"Tour in Disneyland, Hongkong"* and *"Christmas day in Hongkong"*, may have high probabilities on the entries that stand for the topics such as *"amusement parks"* and *"Hongkong"*. By computing the similarity of the topic distribution vectors, we can find the similarity between the corresponding tourists and packages. There are also many other benefits of the TAST model, for example, we can learn the popular topics in each season and find the popular landscapes for each topic.

TABLE 2
Area Segmentation Result

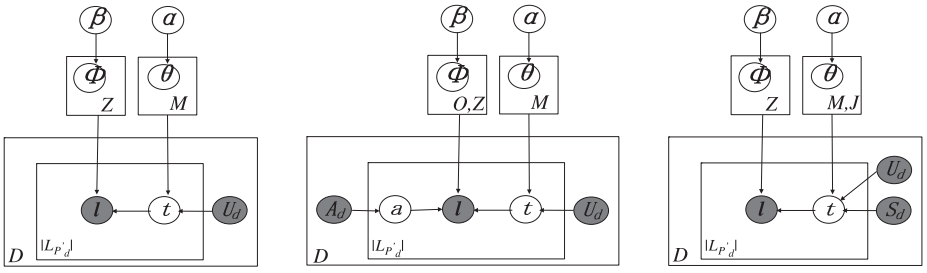
Area	Provinces/Countries	#Landscapes
SC	Guangdong,Guangxi,Macau,Yunnan, Hong Kong,Fujian,Hainan	509
CC	Jiangxi,Guizhou,Sichuan,Hunan,Zhejiang, Jiangsu,Shanghai,Chongqing,Hubei,Anhui	149
NC	Shaanxi,Henan,Heilongjiang,Jilin,Liaoning, Beijing,Tianjing,Shanxi,Shandong,Xinjiang	95
EA	Japan, South Korea	95
SA	Singapore,Malaysia,Thailand,Brunei	118
OC	Australia, New Zealand	55
NA	USA	44

3.3 Area/Seasons Segmentation

There are two extremes for the coverage of each area A_i and each season S_i: we can view the whole earth as an area and the entire year as a season, or we can view each landscape itself as an area and each month as a different season. However, the first extreme is too coarse to capture the spatial-temporal autocorrelations, and we will face the overfitting issue for the second extreme and the Gibbs sampling will be difficult to converge.

To this end, we divide the entire location space in our data set into seven big areas (shown in Table 2) according to the travel area segmentations provided by the travel company, which are South China (SC), Center China (CC), North China (NC), East Asia (EA), Southeast Asia (SA), Oceania (OC), and North America (NA), respectively. To make more reasonable season splitting, we assume that most packages are seasonal, and we use an information gain-based method [12] to get the season splits. The information entropy of the season S^P is $\mathrm{Ent}(S^P) = -\sum_{i=1}^{|S^P|} p_i \log(p_i)$, where $|S^P|$ is the number of different packages in S^P and p_i is the proportion of package P_i in this season. Initially, the entire year is viewed as a big season and then we partition it into several seasons recursively. In each iteration, we use the weighted average entropy (WAE) to find the best split:

2. The price factor will be considered later.



(a) The TT model. (b) The TAT model. (c) The TST model.

Fig. 4. The three related topic models.

$$\mathrm{WAE}(i;S^P)=\frac{|S_1^P(i)|}{|S^P|}\mathrm{Ent}(S_1^P(i))+\frac{|S_2^P(i)|}{|S^P|}\mathrm{Ent}(S_2^P(i)),$$

where $S_1^P(i)$ and $S_2^P(i)$ are two subseasons of season S^P when being splitted at the ith month. The best split month induces a maximum information gain given by $\triangle E(i)$ which is equal to $\mathrm{Ent}(S^P)-\mathrm{WAE}(i;S^P)$.

3.4 Related Topic Models

While the generation processes in TAST are similar to those in the text modeling problems for both documents [5], articles [36] and emails [29], the TAST model is quite different from these traditional ones (e.g., LDA, AT, and ART models). The TAST model has a crucial enhancement by considering the intrinsic features (i.e., location, travel seasons) of the landscapes, and, thus, it can effectively capture the spatial-temporal autocorrelations among landscapes. The benefit is that the TAST model can describe the travel package and the tourist interests more precisely, because the nearby landscapes or the landscapes preferred by the same tourists tend to have the same topic. In addition, the text modeling has the assumption that the words in an email/article are generated by multiple authors, while we assume that the landscapes in the package are generated for the specific tourist of this travel log. Therefore, each single text is considered only once in the text models. However, each package may appear many times in the TAST model according to their records in the travel logs.

Indeed, as shown in Fig. 4, there are three related topic models. The first one (see Fig. 4a) is the tourist topic (TT) model, which does not consider the travel area and travel season factors. The second one (see Fig. 4b) is the tourist-area topic (TAT) model, which only considers the travel area. The third one (see Fig. 4c) is the tourist-season topic (TST) model, which only considers the travel season. All these methods can also be used for package and tourist representation. Finally, note that the graphical representations of TT and TST are similar to the AT model [36] and ART model [29], respectively. However, their differences have been discussed.

4 Cocktail Recommendation Approach

In this section, we propose a cocktail approach on personalized travel package recommendation based on the TAST model, which follows a hybrid recommendation strategy [6] and has the ability to combine many possible constraints that exist in the real-world scenarios. Specifically, we first use the output topic distributions of TAST to find the seasonal nearest neighbors for each tourist, and collaborative filtering will be used for ranking the candidate packages. Next, new packages are added into the candidate list by computing similarity with the candidate packages generated previously. Finally, we use collaborative pricing to predict the possible price distribution of each tourist and reorder the packages. After removing the packages which are no longer active, we will have the final recommendation list.

Fig. 5 illustrates the framework of the proposed cocktail approach, and each step of this approach is introduced in the following sections. We should note that, the major computation cost for this approach is the inference of the TAST model. As the increase of travel records, the computation cost will increase. However, since the topics of each landscape evolves very slowly, we can update the inference process periodically offline in real-world applications. At the end of this section, we will describe many similar cocktail recommendation strategies based on the related topic models of TAST.

4.1 Seasonal Collaborative Filtering for Tourists

In this section, we describe the method for generating the personalized candidate package set for each tourist by the collaborating filtering method. After we have obtained the topic distribution of each tourist and package by the TAST model, we can compute the similarity between each tourist by their topic distribution similarities.

Intuitively, based on the idea of collaborative filtering, for a given user, we recommend the items that are preferred by the users who have similar tastes with her. However, as

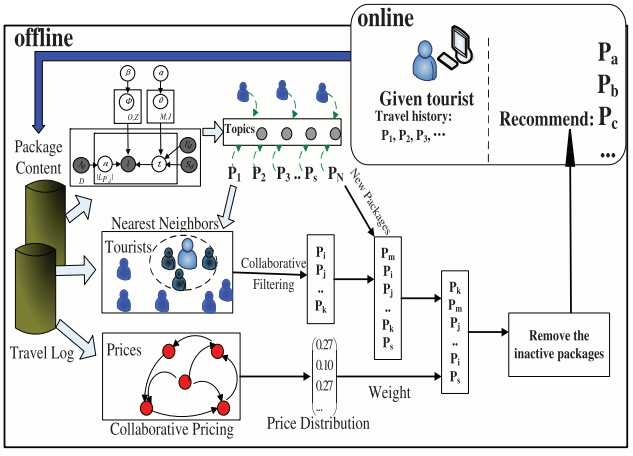


Fig. 5. The cocktail recommendation approach.

we explained previously, the package recommendation is more complex than the traditional ones. For example, if we make recommendations for tourists in winter, it is inappropriate to recommend *"Maple Leaf Adventures."* In other words, for a given tourist, we should recommend the packages that are enjoyed by other tourists at the specific season. Indeed, we have obtained the seasonal topic distribution for each tourist from the TAST model. Multiple methods can be used to compute these similarities, such as matrix factorization [22], [23] and graphical distances [13]. Alternatively, a simple but effective way is to use the *correlation coefficient* [31], and the similarity between tourist U_m and U_n in season S_j can be computed by

$$Sim_{S_j}(U_m, U_n) = \frac{\sum_{k=1}^{Z}(\theta_{mjk} - \bar{\theta}_{mj})(\theta_{njk} - \bar{\theta}_{nj})}{\sqrt{\sum_{k=1}^{Z}(\theta_{mjk} - \bar{\theta}_{mj})^2}\sqrt{\sum_{k=1}^{Z}(\theta_{njk} - \bar{\theta}_{nj})^2}}, \tag{3}$$

where $\bar{\theta}_{mj}$ is the average topic probability for the tourist-season pair (U_m, S_j),[3] For a given tourist, we can find his/her nearest neighbors by ranking their similarity values. Thus, the packages, favored by these neighbors but have not been traveled by the given tourist, can be selected as candidate packages which form a rough recommendation list, and they are ranked by the probabilities computed by the collaborative filtering.

4.2 New Package Problem

In recommender systems, there is a cold-start problem, i.e., it is difficult to recommend new items. As we have explored in Section 2, travel packages often have a life cycle and new packages are usually created. Meanwhile, most of the landscapes will keep in use, which means nearly all the new packages are totally or partially composed by the existing landscapes. Let us take the year of 2010 as an example. There are 65 new packages in the data and only 2 of them are composed completely by new landscapes. Thus, for most of the new packages P^{new}, their topic distributions can be estimated by the topics of their landscapes:

$$\vartheta_{ij}^{P^{new}} = \frac{\alpha_j + \sum_{l \in P_i^{new}} o_{lj}}{\sum_{k=1}^{Z}(\alpha_k + \sum_{l \in P_i^{new}} o_{lk})}, \tag{4}$$

where o_{lj} is the number of times that landscape l is assigned to topic T_j in the travel logs, and the seasonal topic distribution of the new packages can be computed in the similar way. The following question is how to recommend new packages. One way to address this issue is to recommend the new packages that are similar to the ones already traveled by the given tourist (i.e., via the content-based method). However, if the recommender systems just deal with the current interest of the given tourist, we will suffer from the overspecialization problem [2]. Thus, we propose to compute the similarity between the new package and the given number (e.g., 10) of candidate packages in the top of the recommendation list. The new packages which are similar to the candidate packages are added into the recommendation list and their ranks in the list based on the average probabilities of the similar candidate packages. It is expected that this method can not only deal with the cold-start problem but also avoid the overspecialization problem. Please note that, in real applications, new travel package recommendation list can be separated from the general list. However, in this paper, for better illustration and evaluation, we insert the new packages into the general recommendation list as an alternative.

Since there is no effective method to learn the topics of the new packages whose landscapes are not included in the training set, we can use the topic distributions of their located areas on the given travel season as an estimation. Luckily, there are few such packages.

4.3 Collaborative Pricing

In this section, we present the method to consider the price constraint for developing a more personalized package recommender system. The price of travel packages may vary from \$20 to more than 3,000, so the price factor influences the decision of tourists. Along this line, we propose a collaborative pricing method in which we first divide the prices into different segments. Then, we propose to use the Markov forecasting model to predict the next possible price range for a given tourist.

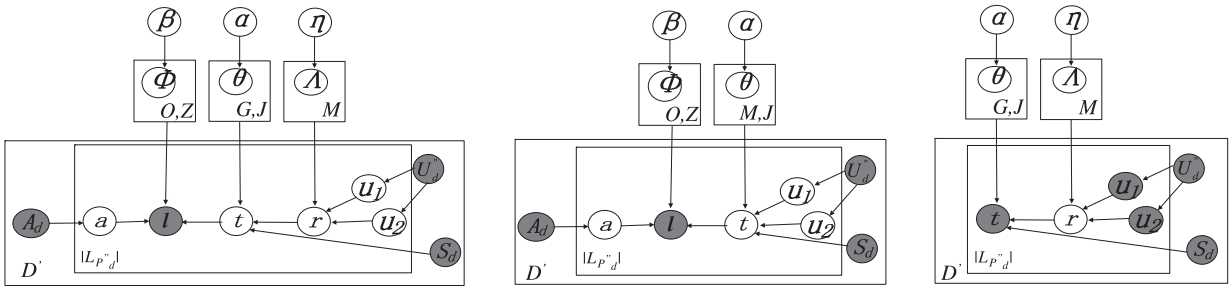
In the first phase, we divide the prices of the packages based on the variance of prices in the travel logs, and the method is similar to the one used in [46]. We first sort the prices of the travel logs, and then partition the sorted list PL into several sublists in a binary-recursive way. In each iteration, we first compute the variance of all prices in the list. Later, the best split price having the minimal weighted average variance (WAV) defined as

$$\mathrm{WAV}(i; PL) = \frac{|PL_1(i)|}{|PL|}\mathrm{Var}(PL_1(i)) + \frac{|PL_2(i)|}{|PL|}\mathrm{Var}(PL_2(i)),$$

where $PL_1(i)$ and $PL_2(i)$ are two sublists of PL split at the ith element and Var represents the variance. This best split price leads to a maximum decrease of $\triangle V(i)$, which is equal to $\mathrm{Var}(PL) - \mathrm{WAV}(i; PL)$.

In the second phase, we mark each price segment as a price state and compute the transition probabilities between them. Specifically, at first, if a tourist used a package with price state a before traveling a package with price state b, then the weight of the edge from a to b will plus 1. After summing up the weights from all the tourists, we normalize them into transition probabilities, and all the transition probabilities compose a state transition matrix. From the current price state of a given tourist (i.e., the current price distribution normalized from his/her previous travel records), we predict the next possible price state by the one-step Markov forecasting model based on random walk. Finally, we obtain the predicted probability distribution of the given tourist on each state, and use these probabilities as weights to multiply the probabilities of the candidate packages in the rough recommendation list so as to reorder these packages. After removing the packages which are no longer active, we have the final recommendation list.

3. If tourist U_m has never traveled in season S_j, then her total topic distribution ϑ_m^U is used as an alternative throughout this paper.



(a) The TRAST model. (b) The $TRAST_1$ model. (c) The $TRAST_2$ model.

Fig. 6. The TRAST model and its two submodels.

4.4 Related Cocktail Recommendations

The previous cocktail recommendation approach (Cocktail) is mainly based on the TAST model and the collaborative filtering method. Indeed, another possible cocktail approach is the content-based cocktail, and in the following, we call this method TASTContent. The main difference between TASTContent and Cocktail is that in TASTContent the content similarity between packages and tourists are used for ranking packages instead of using collaborative filtering. Since TASTContent can only capture the existing travel interests of the tourists, thus it may also suffer from the overspecialization problem.

As there are many related topic models for the TAST model, it is also possible to design the similar cocktail recommendation approaches based on these models. Actually, it is quite straightforward to replace the TAST model by TT, TAT, and TST models in the cocktail approach to get the new recommendations. For example, in the experimental section, the notation TTER stands for the cocktail approach that is based on the TT model.

In Cocktail, we use the price factor as an external constraint to measure package ranks. To some extent, the package prices may also directly influence the interests of the tourists. Thus, it can be included in the topic model representation. If we replace the season token S_d in Fig. 3 by (S_d, C_d) pair, where C_d is the price segment of this package log, and update the previous $4W$ assumptions, the price factor can be well incorporated into the topic model. In this way, the topic preference of the packages in each price segment can also be inferred. What's more, this topic model shares the same inference process with the TAST model, and in the following, we call the cocktail recommendation approach based on this model as Cocktail-.

In summary, both Cocktail and the above related approaches follow the idea of hybrid recommendations, which exploit multiple recommendation techniques, such as collaborative filtering and content-based approaches, for the best performances. Indeed, hybrid recommender systems are usually more practical and have been widely used [6], [24]. For instance, seven different types of hybrid recommendation techniques have been discussed in [6]. In fact, the cocktail recommendation is a combined exploitation of several hybrid approaches. Specifically, the seasonal collaborative filtering based on topic modeling is a *"Feature Augmentation"*, where the new features of latent topics are generated as the better input to enhance the existing algorithm. Second, the insertion of new packages is a *"Mixed"* strategy, where recommendations from different sources are combined. At last, the collaborative pricing is similar to a *"Cascade"* strategy, where the secondary recommender refines the decisions made by a stronger one.

5 The TRAST Model

In this section, we extend the current TAST model and propose a novel tourist-relation-area-season topic model to formulate the tourist relationships in a travel group.

In the TAST model, we do not consider the information of the travel group. However, as noted in Section 2, each package has usually been used by many groups of tourists, and the tourists belong to different travel groups. Thus, if two tourists have taken the same package but in different travel groups, we can only say these two tourists have the same travel interest, but we cannot conclude that they share the same travel profile. However, if these two tourists are in the same group, they may share some common travel traits, such as similar cultural interests and holiday patterns. In the future, they may also want to travel together. Also, they may be family and always travel together during the holiday season. In this paper, we use the notation *relationship* to measure these commonalities and connections in tourists' travel profiles. Please also note that there are multiple tourist relationships simultaneously.

Based on the above understanding, we incorporate into the TAST model a new set of variables, with each entry indicating one relationship, and we consider the tourist relationships in each travel group. This novel topic model is named as the TRAST model, as shown in Fig. 6a, where each tourist has a multinomial distribution over G relationships, and each relationship has a multinomial distribution over Z topics. Other assumptions are similar to those in the TAST model. However, in the TRAST model, the purchases of the tourists in each travel group are summed up as one single expense record and, thus, it has more complex generative process. We can understand this process by a simple example. Assume that two selected tourists in a travel group (U''_d) are u_1 and u_2, who are young and dating with each other. Now, they decide to travel in winter (S_d) and the destination is North America (A_d). To generate a travel landscape (l), we first extract a relationship (r, e.g., lover), and then find a topic (t) for lovers to travel in the winter (e.g., skiing). Finally, based on this skiing topic and the selected travel area (e.g., Northeast America), we draw a landscape (e.g., Stowe, Vermont).

Thus, in the TRAST model, the notation $U_d^{''}$ stands for a group of tourists and $P_d^{''}$ is the corresponding package ID for this travel group log. θ and Λ correspond to the topic distribution and relationship distribution specific to each relationship-season pair and tourist, respectively, where η is a new hyperparameter. The marginal distribution of the travel group set $P^{''}$ can be computed as

$$
\begin{aligned}
&p(P^{''}|\alpha,\beta,\eta,U,S,A)\\
&= \iiint \prod_{i=1}^{M} p(\Lambda_i|\eta) \prod_{i=1}^{G}\prod_{j=1}^{J} p(\theta_{ij}|\alpha) \prod_{i=1}^{O}\prod_{j=1}^{Z} p(\phi_{ij}|\beta) \prod_{d=1}^{D^{'}} \prod_{i=1}^{|L_{P_d^{''}}|}\\
&\left(p(u_1,u_2|U_d^{''}) \sum_{r_{di}=1}^{M}\left(p(r_{di}|u_1,u_2) \sum_{t_{di}=1}^{Z}\left(p(t_{di}|\theta_{r_{di}S_d})\right.\right.\right.\\
&\left.\left.\left.\sum_{a_{di}\in A_d}(p(a_{di}|A_d)p(l_{di}|\phi_{a_{di}t_{di}}))\right)\right)\right) d\phi d\theta d\eta.
\end{aligned}
$$

To perform the inference, the Gibbs sampling formulae can be derived in a similar way as the TAST model, but the sampling procedure at each iteration is significantly more complex. To make inference more efficient and easier for understanding, we instead perform it in two distinct parts. Here, we follow the strategy that is used in [29]. We first split TRAST model into two submodels, as shown in Figs. 6b and 6c. The first submodel TRAST_1 is just like the TAST model, except for the two tourists are latent factors and some of the notations are with different meanings here. By this model, we use a sample to obtain topic assignments and tourist pair assignments for each landscape token. Then, in the second submodel TRAST_2, we treat topics and tourist pairs as known, and the goal is to obtain relationship assignments. In the following, let us introduce the inference of these two models, one by one.

If we directly transfer the results that we get from the TAST model to assign a topic for each landscape token in the TRAST_1 model, we need to compute $n_{(u_1,u_2)st}$ for each (u_1,u_2) pair, which is the number of landscape tokens that are assigned to topic t, and have been cotraveled by tourists (u_1,u_2) in season s. In this way, we have to compute and store each $n_{(u_1,u_2)st}$, an entry in an $M*M*J*Z$ matrix. Thus, the cost will be too expensive (actually, most of the entries should be 0). Instead, we use the following strategy as a simulation:

$$
\begin{aligned}
&p(a_{d_i},t_{d_i},(u_{di_1},u_{di_2})|...)\\
&\propto \frac{\alpha_{t_{di}}+n_{u_{di_1}s_dt_{di}}+n_{u_{di_2}s_dt_{di}}-1}{\Sigma_{k=1}^{Z}\left(\alpha_k+n_{u_{di_1}s_dk}+n_{u_{di_2}s_dk}\right)-1} \frac{\beta_{l_{di}}+m_{a_{di}t_{di}l_{di}}-1}{\Sigma_{k=1}^{|A_i|}\left(\beta_{l_{di}}+m_{a_{di}t_{di}k}\right)-1},
\end{aligned}
\tag{5}
$$

where "..." refers to all the known information such as the area $(\mathbf{a}_{\neg d_i})$, topic $(\mathbf{t}_{\neg d_i})$ and tourist pair $((u1,u2)_{\neg d_i})$ information of other landscape tokens, and the hyperparameters α, and β. By the above equation, we only have to keep a $M*J*Z$ matrix for storing each n_{ust}.

We can see that the TRAST_2 model is similar to the TST model (see Fig. 4c), except for the location of S_d and the pair of tourists. Similar to the inference of the TRAST_1 model, when inferring this model, for each relationship assignment, we use the following equation:

$$
p(r_{d_i}|\ldots) \propto \frac{\eta_{r_{di}}+n_{u_1r_{di}}+n_{u_2r_{di}}-1}{\Sigma_{k=1}^{G}(\eta_k+n_{u_1r_k}+n_{u_2r_k})-1} \frac{\alpha_{t_{di}}+m_{r_{di}S_dt_{di}}-1}{\Sigma_{t=1}^{Z}(\alpha_t+m_{r_{di}S_dt})-1}. \tag{6}
$$

After Gibbs sampling, each tourist's travel relationship preference can be estimated by the following equation, and each entry of θ and ϕ can be computed similarly

$$
\hat{\Lambda_{ir}} = \frac{\eta_t+n_{ir}}{\Sigma_{k=1}^{G}(\eta_k+n_{ik})}. \tag{7}
$$

Actually, this TRAST model can be easily extended for computing relationships among many more tourists. However, the computation cost will also go up. To simplify the problem, in this paper, each time we only consider two tourists in a travel group as a tourist pair for mining their relationships. By this TRAST model, all the tourists' travel preferences are represented by relationship distributions. For a set of tourists, who want to travel the same package, we can use their relationship distributions as features to cluster them, so as to put them into different travel groups. Thus, in this scenario, many clustering methods can be adopted. Since choosing clustering algorithm is beyond the scope of this paper, in the experiments, we refer to K-means [28], one of the most popular clustering algorithms.

Thus, the TRAST model can be used as an assessment for travel group automatic formation. Indeed, in real applications, when generating a travel group, some more external constraints, such as tourists' travel date requirements, the travel company's travel group schedule should also be considered. Please note that, it is possible to use the topics mined by TRAST_1 to represent the latent relationships directly. However, in this way, the topics will represent both landscape topics and latent relationships, it would be hard for interpretation.

6 Experimental Results

In this section, we evaluate the performances of the proposed models on real-world data, and some of previous results [25] are omitted due to the space limit. Specifically, we demonstrate:

1. the results of the season splitting and price segmentation,
2. the understanding of the extracted topics,
3. a recommendation performance comparison between Cocktail and benchmark methods,
4. the evaluation of the TRAST model, and
5. a brief discussion on recommendations for travel groups.

6.1 The Experimental Setup

The data set was divided into a training set and a test set. Specifically, the last expense record of each tourist in the year of 2010 was chosen to be a part of the test set, and the remaining records were used for training. The detailed information is described in Table 3.[4] Note that there are 65 new packages traveled by 269 tourists in

4. Since the data is very sparse and to ensure that each method can get a meaningful result, we choose a comparably small test set.

TABLE 3
The Description of the Training and Test Data

Data Split	#Tourists	#Packages	#Landscapes	#Records	#Groups
Training set	5,211	843	1,054	22,201	7,083
Test set	1,150	908	1,065	1,150	666

the test set. However, only two of these packages are composed completely by new landscapes, and there are 11 new landscapes.

Benchmark methods. To compare the fitness of the TAST model, we compare it with three related models: the TAT model, the TST model, and the TT model, which do not take the season, area, and both season and area factors into consideration, respectively. The perplexity (an evaluation metric for measuring the goodness of fit of a model [29]) comparison result illustrated in [25] shows that TAST model has significantly better predictive power than three other models.

For the recommendation accuracies of the Cocktail approach, we compare it with the following benchmarks:

- Three methods based on topic models including TTER, TASTContent and Cocktail- as described in Section 4.4.
- A content-based recommendation (SContent) based on cotraveled landscapes, following in [27, Eqs. (3.1)-(3.4)].
- For the memory-based collaborative filtering, we implemented the user-based collaborative filtering method (UCF) [31].
- For the model-based collaborative filtering, we chose binary SVD (BSVD) [24].
- Since UCF and BSVD only use the package-level information, to do a fair comparison, we implemented two similar methods based on landscapes (i.e., LUCF, LBSVD).
- One graph-based algorithm, LItemRank [16], where a landscape correlation graph is constructed, and the packages are ranked by the expected average steady-state probabilities on their landscapes.

In the following, we choose the fixed Dirichlet distributions, and these settings are widely used in the existing works [18], [29]. For instance, we set $\beta = 0.1$ and $\alpha = 50/Z$ for the TAST model.

6.2 Season Splitting and Price Segmentation

In this section, we present the results of season splitting and price segmentation as shown in Fig. 7. For better illustration, in Fig. 7a, we only show the travel logs with prices lower than \$1,500. In the figure, different price segments are represented with different grayscale settings, and seasons are split by the dashed lines among months. In total, we have four seasons (i.e., spring, summer, fall, and winter), and five price segments (i.e., very low, low, medium, high, and very high). Since almost all the tourists in the data are from South China, this season splitting has well captured the climatic features there. Another interesting observation is that the peak times for travel in China include February (around the Spring Festival), July and August (the summer for students) and the beginning of October (National Day holiday).

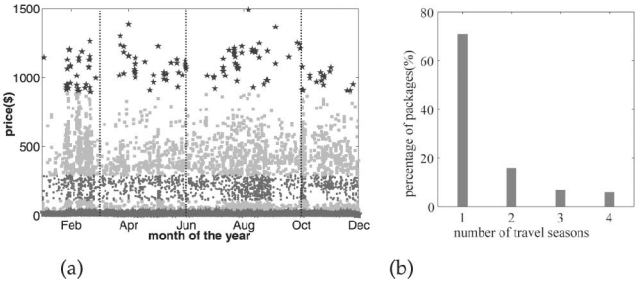


Fig. 7. Season splitting and price segmentation.

Fig. 7b describes the relationship between the percentage of the travel packages and the number of scheduled travel seasons. In Fig. 7b, we can see that most of the packages are only traveled in one season during a year, and less than 6 percent packages are scheduled in the entire year. At last, note that we do not give the illustration of relationship between each travel package and the number of its located areas. The reason is that almost all the packages in the data located in only one of the seven travel areas. These statistical results reflect the fact that landscapes in most packages have spatial-temporal autocorrelations, and the travel area and travel season segmentation methods are reasonable and effective.

6.3 Understanding of Topics

To understand the latent topics extracted by TAST, we focus on studying the relationships between topics and their landscapes'/packages' intrinsic characteristics.

In [25], we have demonstrated that TAST can capture the spatial-temporal correlations among landscapes, and these landscapes, which are close to each other or with similar travel seasons, can be discovered. Meanwhile, the TAST model retains the good quality of the traditional topic models for capturing the relationships between landscapes locating in different areas and has no special travel season preference. Similarly, the topic distributions on each package can be also computed, and Table 4 illustrates many packages with highest probabilities from eight topics identified by the TAST model ($Z = 50$) with the price factor being considered (as illustrated in Section 4.4). Based on the price-spatial-temporal correlations of packages (for many interpretations, there may contain some noise), all the topics can now be classified into eight types, which are noted from 1-1-1 (packages have price, spatial and temporal correlations) to 0-0-0 (packages have none of these correlations). Another interesting observation is that, the top travel packages in many topics are actually quite similar with each other, even though they are with different package IDs. For example, all the packages in topic 43 are about the Kunming-Dali-Lijiang tour. This finding once again demonstrates that, in addition to capture the intrinsic characteristics of the travel data, the TAST model still holds the capability of traditional models, such as the property of clustering documents (packages) [5].

In addition, we show the Pearson correlations of the topic distributions for different prices/areas/seasons in

TABLE 4
An Illustration of Several Topics with Their Travel Packages Having Different Price-Spatial-Temporal Characteristics

Topic 20 (1-1-1)				Topic 43 (1-1-0)			
Package ID, Description	Price	Area	Seasons	Package ID, Description	Price	Area	Seasons
181, 2 days luxury tour in Disneyland	Medium	SC	Spr/Sum	253, 6 days trip to Kunming-Dali-Lijiang (Round trip flight)	High	SC	Entire Year
54, 2 days special tour in Disneyland	Medium	SC	Sum	240, 6 days food discovery trip to Kunming-Dali-Lijiang (Round trip flight)	High	SC	Win
39, 2 days tour in Disneyland	Medium	SC	Sum	359, 9 days trip to Kunming-Dali-Lijiang (double-sleeper)	Medium	SC	Fal
297, 2 days luxury tour in Disneyland/ Hollywood, Hongkong(by ship)	Medium	SC	Spr/Sum	637, 6 days trip to Kunming-Dali-Lijiang (Round trip flight, and membership-only)	High	SC	Fal
13, one day special tour in Hongkong	Low	SC	Spr/Sum	155, 6 days trip to Kunming-Dali-Lijiang (Round trip flight)	High	SC	Sum/Fal
Topic 46 (1-0-1)				**Topic 8 (1-0-0)**			
Package ID, Description	Price	Area	Seasons	Package ID, Description	Price	Area	Seasons
238, 5 days exciting trip to Phuket/ PP islands, Thailand	High	SA	Sum/Fal	55, 5 days tour in Singapore-Malaysia	High	SA	Entire Year
152, 5 days tour in Zhuhai-Beijing	High	NC	Spr/Sum	152, 5 days tour in Zhuhai-Beijing	High	NC	Spr/Sum
11, 5 days tour in Seoul, Jeju island in Korea	High	EA	Spr/Sum	67, 5 days tour in East China	High	CC	Win
53, One day food discovery tour in JiangXin/Macau	VLow	SC	Spr	164, 6 days luxury tour in Honshu Japan	VHigh	EA	Entire Year
291, 6 days tour in Beijing (Round trip flight)	High	NC	Spr/Sum	20, One day food discovery tour in Heshan/Macau	VLow	SC	Entire Year
Topic 36 (0-1-1)				**Topic 17 (0-1-0)**			
Package ID, Description	Price	Area	Seasons	Package ID, Description	Price	Area	Seasons
83, 2 days special travel in Hongkong	Low	SC	Sum	158, 3 days cultural travel in Chaozhou/Shantou/Xiamen	Medium	SC	Win/Spr
13, One day special tour in Hongkong	VLow	SC	Sum/Spr	79, Christmas day in HongKong	VLow	SC	Fal
79, Christmas day in HongKong	VLow	SC	Fal	270, 3 days tour in Chaozhou/ Shantou/Xiamen(2 nights in Xiamen)	Medium	SC	Spr
611, 2 days travel in Hongkong	Medium	SC	Sum/Fal	220, 3 days special tour in Chaozhou/Shantou/Xiamen	Low	SC	Fal
252, 5 days tour in Bangkok-Pattaya	High	SA	Sum	4, 5 days tour in Zhengzhou-Luoyang-Xi'an (Round trip flight)	High	NC	Entire Year
Topic 1 (0-0-1)				**Topic 30 (0-0-0)**			
Package ID, Description	Price	Area	Seasons	Package ID, Description	Price	Area	Seasons
15, 1 day surfing experience in Yangjiang	VLow	SC	Spr/Sum	32, One day food discovery in Jianmen/Macau	VLow	SC	Entire Year
55, 5 days tour in Singapore-Malaysia	High	SA	Spr	88, 5 days Yangtze gorges tour(double-sleeper)	Medium	CC	Spr
218, One day experience in Amusement park,Yangjiang	VLow	SC	Spr/Sum	8, 2 days Middle Autumn tour in Guangdong	Low	SC	Fal
38, 2 days food/constructions/beach discovery in Yangjiang	Low	SC	Sum	41, One day food discovery tour in Kaiping/Enping/Macau	VLow	SC	Entire Year
152, 5 days tour in Zhuhai-Beijing	High	NC	Spr/Sum	100, 7 days Sakura Adventures in Nagoya-Hokkaido-Honshu	VHigh	EA	Spr

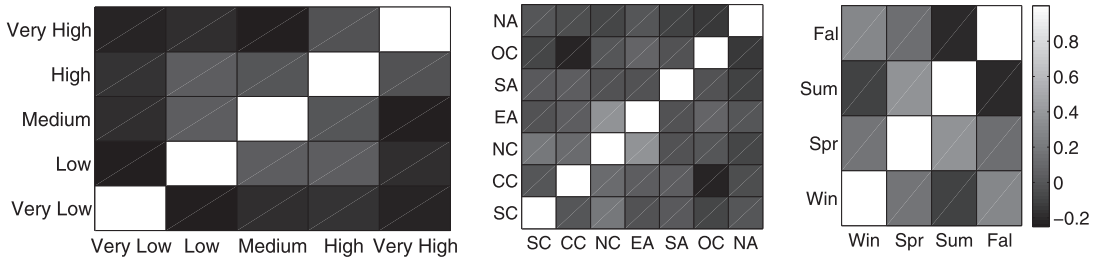


Fig. 8. The correlation of topic distributions between different price ranges (Left)/different areas (Center)/different seasons (Right). Darker shades indicate lower similarity.

Fig. 8 where different prices/areas/seasons are assigned with different topic distributions. From the left matrix, it is very interesting to observe that the topic distribution of the very low price segment and the very high price segment are quite different from three other price ranges. In the center matrix, for most area pairs, there are no obvious topic correlations, except for East Asia (EA) and North China (NC) (for detailed area information, please refer to Table 2), which locate nearby and are with similar latitude. The different types of topic relationships between seasons are more clear as shown in the right matrix, the most different two pairs of seasons are (winter, summer) and (summer, fall), while (summer, spring) have the most similar latent topic distributions.

6.4 Recommendation Performances

Since there are no explicit ratings for validation, we use the ranking accuracy instead. We adopt the widely used degree of agreement (DOA) [26] and Top-K [23] as the evaluation metrics. Also, a simple user study was conducted and volunteers were invited to rate the recommendations. For comparison, we recorded the best performance of each algorithm by tuning their parameters, and we also set some general rules for fair comparison. For instance, for collaborative filtering-based methods, we usually consider the contribution of the nearest neighbors with similarity values larger than 0.

DOA measures the percentage of item pairs ranked in the correct order with respect to all pairs [16]. Let

TABLE 5
A Performance Comparison: DOA (Percent)

Alg.	SContent	UCF	BSVD	LUCF	LBSVD	LItemRank	TTER	TASTContent	Cocktail-	Cocktail
DOA(%)	62.41	69.96	68.77	88.44	87.67	84.76	89.82	80.00	92.44	**92.56**

$NW_{U_i} = P - (F_{U_i} \cup E_{U_i})$ denote the set of packages that do not occur in the training set (F_{U_i}) nor the test set (E_{U_i}) for U_i, and PR_{P_j} denote the predicted rank of package P_j in the recommendation list, and define $check\_order_{U_i}(P_j, P_k)$ as 1 if $PR_{P_j} \geq PR_{P_k}$ otherwise 0. Then the individual DOA for tourist U_i is defined as

$$DOA_{U_i} = \frac{\sum_{j \in E_{U_i}, k \in NW_{U_i}} check\_order_{U_i}(P_j, P_k)}{|E_{U_i}| \times |NW_{U_i}|}.$$

For instance, an ideal(a random) ranking corresponds to a 100 percent (an average 50 percent) DOA, and we use DOA to stand for the average of each individual DOA. Under this metric, the ranking performance of each method is shown in Table 5, where we can see that Cocktail outperforms the benchmark methods. By integrating the price factor into the TAST model, Cocktail- performs nearly as well as Cocktail, and both of them perform better than TTER. Also, the methods that consider landscape information (i.e., LUCF, LBSVD, LItemRank, TTER, TASTContent, Cocktail) usually outperform those do not use such information (i.e., UCF, BSVD). As mentioned previously, it is harder to find the credible nearest neighbor tourists (and latent interests) only based on the cotraveling packages. Furthermore, TASTContent performs better than SContent, and TTER performs better than LUCF and LBSVD, and these demonstrate the effectiveness of modeling latent topics. Meanwhile, unlike watching movies, most of the tourists seldom travel the packages that are similar to the ones that they have already traveled (e.g., have too many identical landscapes), thus content-based methods (i.e., SContent and TASTContent) perform worse than collaborative filterings (e.g., LUCF and Cocktail).

Top-K indicates the recall value of the recommended top-K percent of packages. Since there is only 1 relevant package for each test tourist (i.e., $|E_{U_i}| = 1$), we define $Top - K_{U_i} = \#hit$, where $\#hit$ equals to 1 or 0. Then, the average of individual Top-Ks are used for comparing the performances of the algorithms as shown in Fig. 9. We can see that Cocktail still outperforms other methods and the Top-K result is very similar to the DOA result, except that BSVD/LBSVD are evaluated better now.

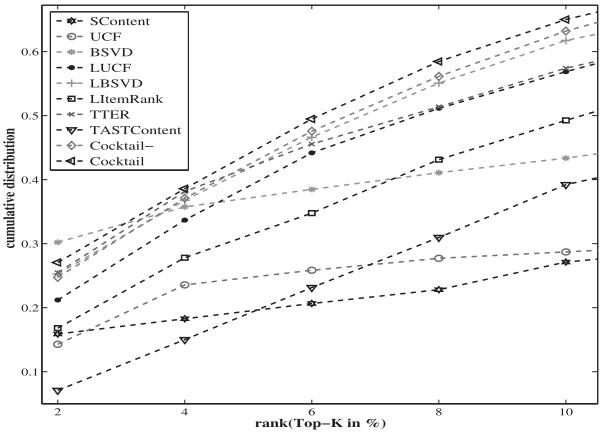


Fig. 9. A performance comparison based on Top-K.

User study. Since it is now impossible for us to directly ask the test tourists to rate the recommendation results, we conducted another type of user study. Specifically, we first gave the package information that one tourist had traveled and the season that he/she was planning a new trip, then we showed the top ranked recommendations from each algorithm (i.e., LUCF, LBSVD, TTER, TASTContent, and Cocktail). Finally, some volunteers were invited to blindly review the recommendations on a five-point Likert scale ranging from 1 (Meaningless) to 5 (Excellent). In total, we collected 2,580 ratings for these five algorithms (i.e., 516 for each) from 17 volunteers (all of them are the undergraduate and graduate students from the University of Science and Technology of China). The final mean ratings and the standard deviations (SD) are shown in Table 6. We can see that the rating for Cocktail is slightly higher than others, and LBSVD outperforms both LUCF and TASTContent. By applying z-test, we find that the differences between the ratings obtained by Cocktail and the other algorithms are statistically significant with $|z| \geq 2.58$ and thus $p \leq 0.01$ (except for the comparison with TTER, where $|z| = 1.53$ and $p = 0.06$). Another interesting observation is that the SD value for TASTContent is extremely high, which means this content-based algorithm makes very distinguishable and controversial recommendations.

In summary, Cocktail performs better than other methods for all the evaluation metrics, and Cocktail-/TTER have the second best performances. Due to the unique characteristics of the travel data, the traditional collaborative filtering methods (UCF and BSVD) do not perform well, and they cannot recommend new packages for tourists. Since different metrics characterize the recommendations from different perspectives, some *"controversial results"* have also been observed (e.g., the different performances of LBSVD). In general, the methods, which consider additional useful information in a proper way, tend to have better performances. During the user study where the users are exposed to many different recommendations simultaneously, we also noticed that it is often hard for them to directly judge two recommendation results from different algorithms. This indicates the issue:

TABLE 6
User Study Ratings

	LUCF	LBSVD	TTER	TASTContent	Cocktail
Mean	3.22	3.30	3.46	3.20	**3.55**
SD	0.74	0.75	0.81	**0.94**	0.76

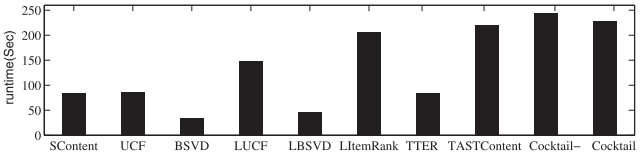


Fig. 10. The runtime results for different algorithms.

TABLE 7
Experimental Results for K-Means Clustering

Features \ Metrics	Cosine		Euclidean distance	
	MI (↑)	VD_n (↓)	MI (↑)	VD_n (↓)
Groups	0.7570	0.4453	0.7659	0.4233
Landscapes	0.7640	0.4727	0.7714	0.4619
Topics	0.7556	0.4227	0.7459	0.4440
Relationships	**0.7972**	**0.4012**	**0.7804**	**0.4161**

the ways of exposing the recommendations and interacting with the users are also very important for successfully deploying a system.

Computational performances. Also, we compare the computational performances of the algorithms. We run all the algorithms on the same platform.[5] Fig. 10 shows the execution time (i.e., the time used for building the model and making final recommendations for all the test tourists). We can see that many algorithms (e.g., LItemRank, TASTContent, Cocktail- and Cocktail) have the similar runtime. Among all the algorithms, BSVD and LBSVD are the most efficient, and Cocktail- has the worst computational performance. Specifically, for the topic model-based methods, TTER does not have to consider the seasonal topic similarities of the tourists, thus it is the most efficient.

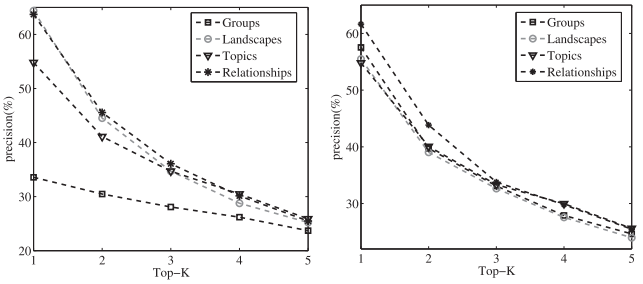
6.5 The Evaluation of the TRAST Model

Since we have little information about tourists, it is hard to interpret the identified relationships. However, we can test the effectiveness of the TRAST model from an alternative perspective; that is, the mined relationships will be used as features to help automatically form travel groups. We conduct two types of experiments. The first experiment is to use K-means clustering for grouping given tourists, and the second one is to find the tourists who would like to travel with given tourist.

To this end, we use 7,083 travel groups to train the TRAST model. For testing, we select 76 packages from the original test set (shown in Table 3) to ensure that each selected package has more than two travel groups. In total, there are 167 travel groups traveled by 570 tourists. In the experiments, we fix the number of topics and relationships to be 100 and 20, and set parameters η, α, and β the same as the TAST model.

For the clustering experiment, given the set of tourists (i.e., objects for clustering) and the number of travel groups (K) of each test package, we run K-means to cluster these tourists into K groups, and here the relationship serves as the feature for clustering. We compare this clustering result with three other clustering results, which are the K-means results by using group logs (i.e., if two tourists often traveled in the same groups, then they will have similar travel preferences), traveled landscapes and topics (mined by TAST model) as features, respectively. Thus, the better the selected features, the better clustering results should be observed. Indeed, K-means clustering validation has been carefully studied before, and we choose two recognized validation measures, MI (mutual information) and VD_n (normalized van Dongen criterion), where MI is a widely used measure and VD_n is the most suitable validation measure identified in [42]. The corresponding experimental results are shown in Table 7. We can see that, regardless of the similarity measures, the K-means results based on relationships always perform much better than the clustering results based on other features for each evaluation metric.

5. For the topic model-based algorithms, we set Gibbs sampling run 100 iterations, since similar results are already observed.



(a) Cosine. (b) Euclid distance.

Fig. 11. The precision results for leave-out-rest (percent).

TABLE 8
The Recall Results for Leave-Out-Rest (Percent)

Features \ Metrics	Cosine	Euclidean distance
Groups	37.10	53.31
Landscapes	59.28	50.27
Topics	53.26	53.07
Relationships	**60.18**	**56.23**

Meanwhile, we evaluate the identified relationships from each tourist's point of view. Specifically, we randomly select a tourist from each travel group, and then we rank all the rest tourists (including the ones from other groups) of this travel package for this tourist (i.e., leave-out-rest). Here, the ranking list is generated based on the candidates' similarities with the given tourist computed by the travel relationship distributions (or cotraveled groups, or landscapes, or topic distributions). Ideally, the tourists who are in the same travel group with the given tourist should appear earlier in the list. To evaluate these ranking lists, we choose "*precision*" and "*recall*" as the metrics, and the corresponding results are shown in Fig. 11 and Table 8. We can see that the ranking lists based on relationships are still better than those based on other features.

From the above analysis, we know that the relationships identified by TRAST can be better used for clustering tourists and help to find the most possible cotravel tourists for a given tourist. Thus, compared to cotraveled groups, landscapes and topics, it is more suitable for travel companies to choose relationships as an assessment for travel group automatic formation.

TABLE 9
Group Recommendation Results: DOA (Percent)

Alg.	LUCF	LBSVD	TTER	Cocktail (Topics)	Cocktail (Relationships)
DOA(%)	90.86	88.77	89.60	**92.29**	92.10

6.6 Recommendation for Travel Groups

The evaluations in previous sections are mainly focused on the individual (personalized) recommendations. Since there are tourists who frequently travel together, it is interesting to know whether the latent variables (e.g., the topics of each individual tourist and the relationships of a travel group) as well as the cocktail approaches are useful for making recommendations to a group of tourists. To this end, we performed an experimental study on group recommendations.

Similar to the evaluation for the personalized recommendation, we recommended for the 666 travel groups existing in the test set (shown in Table 3). Specifically, each recommendation algorithm simply views a group of tourists as an *"individual tourist"* and all the previous travel/expense records of these tourists are used for training, and then generates a single recommendation list for each test group (tourists in this group) using the training set (training groups). According to their performances in Section 6.4, we chose five typical recommendation algorithms for comparison including LUCF, LBSVD, TTER, the two Cocktails based on the topics extracted by the TAST model and based on the relationships extracted by the TRAST model. We chose DOA as the evaluation metric due to its simplicity in interpretation. The experimental results are shown in Table 9, where we can see that Cocktails still outperform other algorithms, and in addition to modeling each individual tourist, the relationships can also be used for making recommendations. Meanwhile, we observe that both LUCF and LBSVD perform much better with more training records comparing to the results in Table 5.

It is worth noting that the differences between group recommendation and individual recommendation are more subtle and complex than we could imagine at the first glance [21]. While the detailed discussion is beyond the scope of this paper, we hope there are more future studies on travel group recommendations.

7 Related Work

In general, the related work can be grouped into two categories. *The first category* has a focus on the recommendation studies in the tourism domain, where some systems have been developed to help tourists and these related work can be divided into two groups.

In the first group, people are focused on the development of intelligent systems for the tourists in the pretravel stage for travel planning [44], information filtering [41], and inspiration [32]. For instance, Yin et al. [44] proposed an automatic trip planning framework by leveraging geotagged photos and textual travel logs. Also, Hao et al. [19] proposed a location-topic model by learning the local and global topics to mine the location-representative knowledge from a large collection of travel logs, and to recommend the travel destinations. Wu et al. [41] designed a system using the multimedia technology to generate the personalized tourism summary. In addition, Ricci et al. [32], [33] described case-based reasoning approaches, Trip@dvice and DieToRecs. By exploiting a set of features for each tourist's specific interaction session, these two approaches address a number of travel issues (e.g., mix-and-match travel planning) and have been successfully used in several websites, such as the visiteurope.com [39]. Finally, by taking the travel cost into the consideration, Ge et al. [14] and Xie et al. [43] provided focused studies of cost-aware tour recommendation.

In the second subgroup, people target on providing more context-aware travel information to the on-tour tourists with mobile devices [35]. For instance, the studies in [1] and [8] aim on the development of mobile tourist guides. Also, Averjanova et al. [4] developed a map-based mobile system that can provide users with some personalized recommendations. Moreover, Carolis et al. [7] used a map for outlining the location and the information of landscapes in a town area. Finally, a more sophisticated on-tour support system, MobyRek, was recently developed by Ricci and Nguyen [34].

In summary, the above systems and algorithms target on helping tourists from different perspectives. However, tourists need system support throughout stages of travel, beginning from pretravel planning through to the final stages of travel [34]. Thus, the real-world travel recommender systems are usually very complicated, and some of the critical gaps, general problems and issues of travel recommendations have been extensively discussed in [17], [32], [33], [38].

The second category includes the research work related to topic models and their applications on recommender systems. Topic models are usually based upon the idea that documents (including messages, emails, etc.) are mixtures of latent topics, where a topic is a probability distribution over words.

Many topic models have been proposed. Among them, the latent Dirichlet allocation (LDA) [5] model possesses fully generative semantics, and thus has been widely studied and extended for many applications. For example, Rosen-Zvi et al. [36] extended LDA to the author-topic (AT) model for computing similarity between authors and the entropy of author output. Based on LDA and AT models, McCallum et al. [29] provided the author-recipient-topic (ART) model for social network analysis. There are also some works, which have successfully applied LDA into the recommendation algorithms. For instance, Chen et al. [9] adopted LDA to model user-community cooccurrences in social networking services, and then made community recommendations. Liu et al. [26] used the LDA model to mine the user latent interests so as to enhance collaborative filtering. Wang and Blei [40] combined the merits of traditional collaborative filtering and probabilistic topic model together to recommend scientific articles.

In summary, topic models (especially LDA)-based methods perform well in the text-related or other recommendation tasks. Furthermore, we can easily integrate heterogeneous data sources into a unified framework by extending current models [29], [36], [44]. Thus, more and more researchers try to exploit topic models for better usage, such as finding the way to combine topic models and matrix factorizations [3], [40].

8 Discussion

Here, we discuss the advantages and limitations of this study. From the experimental results, we can see that the proposed cocktail recommendation approach works very well for predicting the tourists' travel preferences by exploiting the unique characteristics of the travel package data. Also, in this paper, we describe the work in a domain-depended (i.e., travel) way where users are tourists, items are travel packages, and features of items are seasons, areas, and so on. However, it is worth noting that the idea of profiling user/item and the way to explore features and integrate these features in topic modeling should be generally applicable to other recommendation scenarios.

Meanwhile, the cocktail approach has some limitations. First, unlike some intelligent systems [38], the cocktail approach disregards the specific preferences of the tourist while he/she is planning a trip, such as the transportation preference. In other words, we focus on designing the recommendation algorithm to attract the tourists before they make a travel decision rather than providing the travel support in the on-tour stage [32]. Thus, our approach may be only useful in some situations (e.g., email marketing). Also, if we want to deploy this work for real-world services, we have to incorporate more practical functions. Second, there are some limitations with the performance evaluation, which is just based on the ability to recover omitted (hide) test data [20] and a simple user study. For instance, we cannot always attribute a user not traveling a package locating in the top of the recommendation list to a lack of interest of that package [30]; that is, relevant (desirable) travel packages in the test set may be just a small fraction of the entire relevant ones that are actually of interest to each tourist. For real-world applications, more sophisticated online experiments are required.

9 Concluding Remarks

In this paper, we present study on personalized travel package recommendation. Specifically, we first analyzed the unique characteristics of travel packages and developed the TAST model, a Bayesian network for travel package and tourist representation. The TAST model can discover the interests of the tourists and extract the spatial-temporal correlations among landscapes. Then, we exploited the TAST model for developing a cocktail approach on personalized travel package recommendation. This cocktail approach follows a hybrid recommendation strategy and has the ability to combine several constraints existing in the real-world scenario. Furthermore, we extended the TAST model to the TRAST model, which can capture the relationships among tourists in each travel group. Finally, an empirical study was conducted on real-world travel data. Experimental results demonstrate that the TAST model can capture the unique characteristics of the travel packages, the cocktail approach can lead to better performances of travel package recommendation, and the TRAST model can be used as an effective assessment for travel group automatic formation. We hope these encouraging results could lead to many future work.

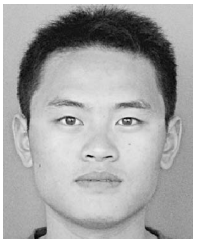
Acknowledgments

This paper was an expanded version of [25], which appeared in the *Proceedings of IEEE 2011 International Conference on Data Mining (ICDM 2011)* as the Best Research Paper. This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), Natural Science Foundation of China (Grant No. 61073110 and 71329201), and the US National Science Foundation (NSF) via grant numbers CCF-1018151 and IIS-1256016. Qi Liu gratefully acknowledges the support of the Fundamental Research Funds for the Central Universities of China, and the Youth Innovation Promotion Association, CAS.

References

[1] G.D. Abowd et al., "Cyber-Guide: A Mobile Context-Aware Tour Guide," *Wireless Networks,* vol. 3, no. 5, pp. 421-433, 1997.
[2] G. Adomavicius and A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. Knowledge and Data Eng.,* vol. 17, no. 6, pp. 734-749, June 2005.
[3] D. Agarwal and B. Chen, "fLDA: Matrix Factorization through Latent Dirichlet Allocation," *Proc. Third ACM Int'l Conf. Web Search and Data Mining (WSDM '10),* pp. 91-100, 2010.
[4] O. Averjanova, F. Ricci, and Q.N. Nguyen, "Map-Based Interaction with a Conversational Mobile Recommender System," *Proc. Second Int'l Conf. Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM '08),* pp. 212-218, 2008.
[5] D.M. Blei, Y.N. Andrew, and I.J. Michael, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, 2003.
[6] R. Burke, "Hybrid Web Recommender Systems," *The Adaptive Web,* vol. 4321, pp. 377-408, 2007.
[7] B.D. Carolis, N. Novielli, V.L. Plantamura, and E. Gentile, "Generating Comparative Descriptions of Places of Interest in the Tourism Domain," *Proc. Third ACM Conf. Recommender Systems (RecSys '09),* pp. 277-280, 2009.
[8] F. Cena et al., "Integrating Heterogeneous Adaptation Techniques to Build a Flexible and Usable Mobile Tourist Guide," *AI Comm.,* vol. 19, no. 4, pp. 369-384, 2006.
[9] W. Chen, J.C. Chu, J. Luan, H. Bai, Y. Wang, and E.Y. Chang, "Collaborative Filtering for Orkut Communities: Discovery of User Latent Behavior," *Proc. ACM 18th Int'l Conf. World Wide Web (WWW '09),* pp. 681-690, 2009.
[10] N.A.C. Cressie, *Statistics for Spatial Data.* Wiley and Sons, 1991.
[11] J. Delgado and R. Davidson, "Knowledge Bases and User Profiling in Travel and Hospitality Recommender Systems," *Proc. ENTER 2002 Conf. (ENTER '02),* pp. 1-16, 2002.
[12] U.M. Fayyad and K.B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI),* pp. 1022-1027, 1993.
[13] F. Fouss et al., "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," *IEEE Trans. Knowledge and Data Eng.,* vol. 19, no. 3, pp. 355-369, Mar. 2007.
[14] Y. Ge et al., "Cost-Aware Travel Tour Recommendation," *Proc. 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '11),* pp. 983-991, 2011.
[15] Y. Ge et al., "An Energy-Efficient Mobile Recommender System," *Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '10),* pp. 899-908, 2010.
[16] M. Gori and A. Pucci, "ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines," *Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07),* pp. 2766-2771, 2007.
[17] U. Gretzel, "Intelligent Systems in Tourism: A Social Science Perspective," *Annals of Tourism Research,* vol. 38, no. 3, pp. 757-779, 2011.
[18] T.L. Griffiths and M. Steyvers, "Finding Scientific Topics," *Proc. Nat'l Academy of Sciences USA,* vol. 101, pp. 5228-5235, 2004.
[19] Q. Hao et al., "Equip Tourists with Knowledge Mined from Travelogues," *Proc. 19th Int'l Conf. World Wide Web (WWW '10),* pp. 401-410, 2010.
[20] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating Collaborative Filtering Recommender Systems," *ACM Trans. Information Systems,* vol. 22, no. 1, pp. 5-53, 2004.
[21] A. Jameson and B. Smyth, "Recommendation to Groups," *The Adaptive Web,* vol. 4321, pp. 596-627, 2007.

[22] Y. Koren and R. Bell, "Advances in Collaborative Filtering," *Recommender Systems Handbook,* chapter 5, pp. 145-186, 2011.
[23] Y. Koren, "Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model," *Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08),* pp. 426-434, 2008.
[24] S. Lai et al., "Hybrid Recommendation Models for Binary User Preference Prediction Problem," *Proc. KDD- Cup 2011 Competition,* 2011.
[25] Q. Liu, Y. Ge, Z. Li, H. Xiong, and E. Chen, "Personalized Travel Package Recommendation," *Proc. IEEE 11th Int'l Conf. Data Mining (ICDM '11),* pp. 407-416, 2011.
[26] Q. Liu, E. Chen, H. Xiong, C. Ding, and J. Chen, "Enhancing Collaborative Filtering by User Interests Expansion via Personalized Ranking," *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics,* vol. 42, no. 1, pp. 218-233, Feb. 2012.
[27] P. Lops, M. Gemmis, and G. Semeraro, "Content-Based Recommender Systems: State of the Art and Trends," *Recommender Systems Handbook,* chapter 3, pp. 73-105, 2010.
[28] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Berkeley Symp. Math. Statistics and Probability (BSMSP),* vol. 1, pp. 281-297, 1967.
[29] A. Mccallum, X. Wang, and A. Corrada-Emmanuel, "Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email," *J. Artificial Intelligence Research,* vol. 30, pp. 249-272, 2007.
[30] R. Pan et al., "One-Class Collaborative Filtering," *Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08),* pp. 502-511, 2008.
[31] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proc. ACM Conf. Computer Supported Cooperative Work (CSCW '94),* pp. 175-186, 1994.
[32] F. Ricci, D. Cavada, N. Mirzadeh, and N. Venturini, "Case-Based Travel Recommendations," *Destination Recommendation Systems: Behavioural Foundations and Applications,* chapter 6, pp. 67-93, 2006.
[33] F. Ricci et al., "DieToRecs: A Case-Based Travel Advisory System," *Destination Recommendation Systems: Behavioural Foundations and Applications,* chapter 14, pp. 227-239, 2006.
[34] F. Ricci and Q. Nguyen, "Mobyrek: A Conversational Recommender System for On-the-Move Travelers," *Destination Recommendation Systems: Behavioural Foundations and Applications,* chapter 17, pp. 281-294, 2006.
[35] F. Ricci, "Mobile Recommender Systems," *Information Technology and Tourism,* vol. 12, no. 3, pp. 205-231, 2011.
[36] M. Rosen-Zvi et al., "The Author-Topic Model for Authors and Documents," *Proc. 20th Conf. Uncertainty in Artificial Intelligence (UAI '04),* pp. 487-494, 2004.
[37] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender Systems-a Case Study," *Proc. ACM WebKDD Workshop,* pp. 82-90, 2000.
[38] S. Staab et al., "Intelligent Systems for Tourism," *IEEE Intelligent Systems,* vol. 17, no. 6, pp. 53-66, Nov. 2002.
[39] A. Venturini and F. Ricci, "Applying Trip@ Dvice Recommendation Technology to www.visiteurope.com," *Proc. 17th European Conf. Frontiers in Artificial Intelligence and Applications,* vol. 141, p. 607, 2006.
[40] C. Wang and D. Blei, "Collaborative Topic Modeling for Recommending Scientific Articles," *Proc. ACM 17th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 448-456, 2011.
[41] X. Wu, J. Li, and S. Neo, "Personalized Multimedia Web Summarizer for Tourist," *Proc. 17th Int'l Conf. World Wide Web (WWW '08),* pp. 1025-1026, 2008.
[42] J. Wu, H. Xiong, and J. Chen, "Adapting the Right Measures for K-Means Clustering," *Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 877-886, 2009.
[43] M. Xie, L.V.S. Lakshmanan, and P.T. Wood, "Breaking Out of the Box of Recommendations: From Items to Packages," *Proc. Fourth ACM Conf. Recommender Systems (RecSys '10),* pp. 151-158, 2010.
[44] H. Yin, X. Lu, C. Wang, N. Yu, and L. Zhang, "Photo2Trip: An Interactive Trip Planning System Based on Geo-Tagged Photos," *Proc. ACM Int'l Conf. Multimedia (MM '10),* pp. 1579-1582, 2010.
[45] Z. Yin et al., "Geographical Topic Discovery and Comparison," *Proc. 20th Int'l Conf. World Wide Web (WWW '11),* pp. 247-256, 2011.
[46] J. Yuan et al., "T-Drive: Driving Directions Based on Taxi Trajectories," *Proc. 18th SIGSPATIAL Int'l Conf. Advances in Geographic Information Systems (GIS '10),* pp. 99-108, 2010.



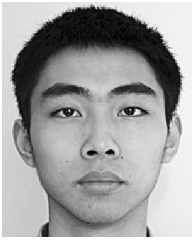
Qi Liu received the BE degree in computer science from Qufu Normal University, China, and the PhD degree from the University of Science and Technology of China (USTC), China. He is currently an associate researcher in the School of Computer Science and Technology at USTC. His major research interests include intelligent data analysis, recommender system, social network and context-aware data mining. He has published several papers in refereed conference proceedings and journals, such as IEEE ICDM'11, ACM SIGKDD'11, IJCAI'13, IEEE ICDM'13, the *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, the *IEEE Transactions on Knowledge and Data Engineering*, the *ACM Transactions on Information Systems*, and the *ACM Transactions on Intelligent Systems and Technology*. He is the recipient of the President's Exceptional Student Award, Chinese Academy of Sciences (CAS), in 2012. He is a member of the China Computer Federation and a member of the Youth Innovation Promotion Association, CAS.

Enhong Chen (SM'07) received the PhD degree from the University of Science and Technology of China (USTC). He is a professor and vice dean of the School of Computer Science and Technology at USTC. His general area of research includes data mining, personalized recommendation systems, and web information processing. He has published more than 100 papers in refereed conferences and journals. His research is supported by the National Natural Science Foundation of China, National High Technology Research and Development Program 863 of China, etc. He is a program committee member of more than 40 international conferences and workshops. He is a senior member of the IEEE.

Hui Xiong (SM'07) received the BE degree from the University of Science and Technology of China (USTC), the MS degree from the National University of Singapore (NUS), and the PhD degree from the University of Minnesota (UMN). He is currently an associate professor and vice chair of the Management Science and Information Systems Department, and the director of Rutgers Center for Information Assurance at the Rutgers, the State University of New Jersey, where he received a two-year early promotion/tenure in 2009, the Rutgers University Board of Trustees Research Fellowship for Scholarly Excellence in 2009, and the ICDM-2011 Best Research Paper Award in 2011. His general area of research is data and knowledge engineering, with a focus on developing effective and efficient data analysis techniques for emerging data intensive applications. He has published prolifically in refereed journals and conference proceedings (3 books, 40+ journal papers, and 60+ conference papers). He is a co-editor-in-chief of *Encyclopedia of GIS* and an associate editor of the *IEEE Transactions on Data and Knowledge Engineering* (*TKDE*) and the *Knowledge and Information Systems* (*KAIS*) journal. He has served regularly on the organization and program committees of numerous conferences, including as a program cochair of the Industrial and Government Track for the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) and a program cochair for the IEEE 2013 International Conference on Data Mining (ICDM). He is a senior member of the ACM and the IEEE, and a member of the ACM SIGKDD.

Yong Ge received the BE degree in information engineering from Xi'an Jiao Tong University, China, in 2005, the MS degree in signal and information processing from the University of Science and Technology of China, Hefei, in 2008, and the PhD degree in Information Technology from Rutgers, The State University of New Jersey in 2013. He is currently an assistant professor at the University of North Carolina at Charlotte. His research interests include data mining and business analytics. He received the ICDM-2011 Best Research Paper Award, Excellence in Academic Research (one per school) at Rutgers Business School in 2013, and the Dissertation Fellowship at Rutgers University in 2012. He has published prolifically in refereed journals and conference proceedings, such as the *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, the *ACM Transactions on Information Systems*, the *ACM Transactions on Knowledge Discovery from Data*, the *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM SIGKDD, SIAM SDM, IEEE ICDM, and ACM RecSys. He has served as a program committee member for the ACM SIGKDD 2013, the International Conference on Web-Age Information Management 2013, and IEEE ICDM 2013. Also he has served as a reviewer for numerous journals, including *TKDE*, *TIST*, *Knowledge and Information Systems (KAIS)*, *Information Science*, and the *IEEE Transactions on Systems, Man, and Cybernetics, Part B*.

Zhongmou Li received the BE degree in computer software from Tsinghua University, Beijing, China, in 2008. He is currently working toward the PhD degree in the Department of Management Science and Information Systems at Rutgers University under the advisory of professor Hui Xiong. His research interests include data mining, financial fraud detection, recommender systems, query clustering, and their applications in real-world business domains. During his PhD study, he has three papers published in the IEEE International Conference on Data Mining. He has also been a reviewer for the *Journal of Knowledge and Information Systems* and an external reviewer for various international conferences, such as KDD, ICDM, CIKM, SDM, and EC.



Xiang Wu received the BE degree in software engineering from Anhui University, Hefei, China, in 2011. He is currently working toward the ME degree in the School of Computer Science and Technology, University of Science and Technology of China. His research interests now include recommender system, community discovery, and web data mining. As a student leader of the ACM-ICPC Association of Anhui University, he has won two bronze medals and one silver medal in the ACM-ICPC Asia Regional Contests.