2011 11th IEEE International Conference on Data Mining

Personalized Travel Package Recommendation

Qi Liu[1,2], Yong Ge[2], Zhongmou Li[2], Enhong Chen[1,*], Hui Xiong[2,*]
[1]*School of Computer Science and Technology, University of Science and Technology of China*
feiniaol@mail.ustc.edu.cn, cheneh@ustc.edu.cn
[2]*MSIS Department, Rutgers Business School, Rutgers University, USA*
yongge@pegasus.rutgers.edu, mosesli@pegasus.rutgers.edu, hxiong@rutgers.edu

***Abstract*—As the worlds of commerce, entertainment, travel, and Internet technology become more inextricably linked, new types of business data become available for creative use and formal analysis. 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 this end, we first analyze the characteristics of the travel packages and develop a Tourist-Area-Season Topic (TAST) model, which can extract the topics conditioned on both the tourists and the intrinsic features (i.e. locations, travel seasons) of the landscapes. Based on this TAST model, we propose a cocktail approach on personalized travel package recommendation. Finally, we evaluate the TAST model and the cocktail approach on real-world travel package data. The 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 methods for travel package recommendation.**

I. Introduction

Tourism has become one of the world's largest industries. Furthermore, according to the forecast by the World Travel & Tourism council, the contribution of tourism to global GDP is expected to rise from 9.1% in 2011 to 9.6% by 2021 [1]. Indeed, with the advancement of time and the improvement of living standards, even an ordinary family can do extended travel very comfortably on a small budget.

As a trend, more and more travel companies, such as Expedia [2], 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 travel packages to satisfy their personalized requirements. On the other side, to get more business and profit, the travel companies have to understand these preferences from different tourists and serve more attractive packages. Therefore, the demand for intelligent travel services, from both tourists and travel companies, is expected to increase dramatically.

Since recommender systems have been successfully applied to enhance the quality of service for customers in a number of fields [2], [14], [20], it is natural direction to develop recommender systems for personalized travel package recommendation. Recommender systems for tourists have been studied before [1], [3], [6], [7]. For instance, the works in [1], [6] target the development of mobile tourist guides. Also, Averjanova et al. have developed a map-based mobile recommender system that can provide users with some personalized recommendations [3]. However, the prior works above are only exploratory in nature, and the problem of leveraging unique features to distinguish personalized travel package recommendations remains pretty much open.

As a matter of fact, 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 cost for a travel is much more expensive than for watching a movie. It is normal for a customer to watch more than one movie each month, while they may only travel one or two times per year. Second, every travel package consists of many landscapes or attractions, and thus has intrinsic complex spatio-temporal relationships. For example, a travel package only includes the landscapes/attractions which are geographically co-located 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 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. However, the values of travel packages can easily depreciate over time, and a tour package usually only lasts for a certain period. The travel companies need to actively create new tour packages to replace the old ones based on the interests of the customers.

To address the challenges mentioned above, in this paper, we propose a cocktail approach on personalized travel package recommendation. Specifically, we first analyze the key characteristics of the travel packages. Along this line, travel time and travel destinations are divided into different seasons and areas. Then, we develop a Tourist-Area-Season Topic (TAST) model, which can extract the topics conditioned on both the tourists and the intrinsic features (i.e. locations, travel seasons) of the landscapes in travel packages. As a result, the TAST model can well represent the content of the travel packages and the interests of the tourists. Based

*Contact Authors.
[1]WTTC, URL:http://www.wttc.org/
[2]Expedia, URL: http://www.expedia.com/

1550-4786/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDM.2011.118

Figure 1. An example of the travel package document, where the landscapes are represented by the words in red.

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 cocktail approach is evaluated on real-world data. The experimental results show that the TAST model can effectively capture the unique characteristics of the travel data and the cocktail approach performs much better than traditional recommender systems.

II. Concepts and Data Description

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

In this paper, we aim to make personalized travel package recommendations for the tourists who have traveled at least once in the existing travel records. In this recommender system, the users are the tourists and the items are the packages. Here, **package** means the comprehensive services provided by a travel company for the tourists based on one or several themes (or **topics**), one travel package usually consists of many **landscapes** located in one or more **areas** as well as some related services such as the transportation, the price, etc. Figure 1 shows an example document for a travel package named "*Niagara Falls Discovery*" from the STA Travel [3]. It includes the topics (tour style), travel days, price, travel area (the northeastern U.S.), the arrangement, and landscapes etc. Note that different packages may include the same landscapes and each landscape can be used for multiple packages. In addition to this, each package has a travel schedule and many packages will be traveled only in a given time (**season**) of the year, which means they have strong seasonal effects. For example, the *"Maple Leaf Adventures"* is only meaningful in fall.

In this paper, we exploit a real world travel data set for building personalized recommender systems for target tourists. This data set is provided by a travel company in China. There are nearly $220,000$ expense records with the travel time stamp between the beginning of 2000 and October 2010. From this data set, we extracted $23,351$ useful

[3]STA Travel, URL:http://www.statravel.com/

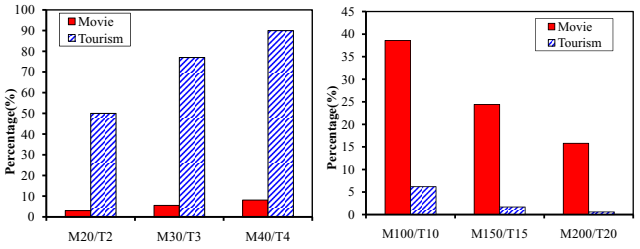


Figure 2. A simple comparison of the data sparseness between the movieLens data and the travel data. (a):The percentage of users/tourists whose co-rating movies/co-traveling packages with their nearest neighbors are no more than (20, 30,40 for the movie users)/(2,3,4 for the tourists). (b): The percentage of users/tourists whose rating movies are more than 100,150,200/whose traveling logs are 10,15,20, respectively.

records from $5,211$ tourists for 908 domestic and international packages to make sure each tourist has traveled at least two different packages (one package can be used for training and another for test). The extracted data contains $1,065$ different landscapes located in 139 cities from 10 countries. On average, each package has 11 different landscapes.

There are some unique characteristics of this travel data. First, it is very sparse. Figure 2 shows a comparison of this travel data with the standard 100K movie recommendation data [4]. The difference can be easily observed. For the comparisons, we chose the movie data 10 times larger than the travel data. In Figure 2(a), we can notice that it is harder to find the credible nearest neighbors for the tourists because there are very few co-traveling packages. Moreover, in Figure 2(b), we can see that nearly 95% of tourists have traveled less than 10 times. On average, each tourist has traveled only 4.4 times and only 0.48% of the entries in the corresponding tourist package matrix are non-zero. This density value is much smaller than those of traditional data sets for recommendation, such as the Netflix [5] data which has the density 1.17%. The extreme sparseness of the travel data raises the challenges for using traditional recommendation techniques, such as the collaborative filtering techniques.

Second, the travel data has much stronger time dependence as shown in Figure 3. Unlike the traditional recommended items, such as movies, and songs, the travel packages often have a life cycle along with the change to the business demand. This means that the package only lasts for a certain period. In contrast, most of the landscapes in these packages 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. From Figure 3, we can observe that the landscapes are more sustainable and more important than the package itself.

Third, each landscape has a geographic location and the right travel seasons. These location and travel season information can be viewed as the intrinsic features of the landscapes. Only the landscapes with similar spatial-

[4]MovieLens, URL:http://www.grouplens.org/node/73
[5]Netflix, URL: http://www.netflixprize.com/

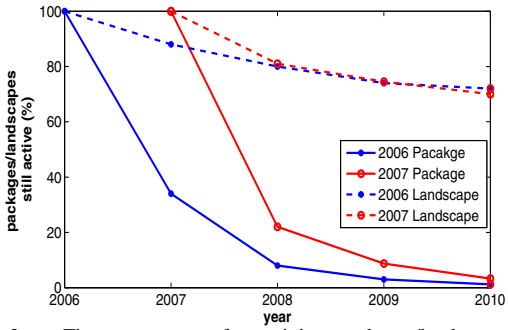


Figure 3. The percentage of remaining packages/landscapes in the following several years after they have been introduced.

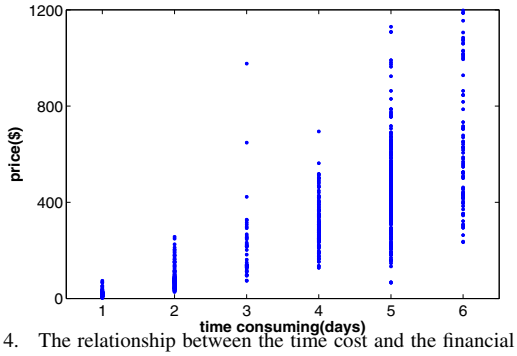


Figure 4. The relationship between the time cost and the financial cost in travel packages.

temporal features are suitable for the same package. This means the landscapes in one package have spatial-temporal auto-correlations, and follow the first law of geography– everything is related to everything else, but the nearby things are more related than distant things [8]. Hence, when making recommendations, we should take the landscapes and their spatial-temporal correlations into consideration so as to describe the tourists and the packages more precisely.

In addition, the price of the travel package can also influence recommendations. Indeed, 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, when developing recommender systems, it is very important to profile the tourists based on their interests as well as the time and the money they can afford. Figure 4 illustrates the relationship between the time and financial cost for each travel package, we can see that the package with a higher price often tends to have more time and vice versa. This means we can use one factor to describe another. In this paper, we only take the price factor into consideration.

Last but not least, unlike the movie data [12], 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 by the package.

In summary, the above analysis shows that the travel data is quite different from the traditional data for recommendation. As a result, it is necessary to develop more suitable approaches for travel package recommendation.

Table I
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
$T = \{T_1, T_2, ..., T_K\}$	the set of topics
$A = \{A_1, A_2, ..., A_O\}$	the set of different areas
$P^{'} = \{P_1^{'}, P_2^{'}, ..., P_D^{'}\}$	the set of package logs
$L_{A_i} = \{L_{A_i 1}, ..., L_{A_{i\|A_i\|}}\}$	landscape set for area A_i
$L_{P_i^{'}} = \{L_{P_i^{'} 1}, ..., L_{P_i^{'}\|P_i^{'}\|}\}$	landscape set for package log $P_i^{'}$

III. THE TAST TOPIC MODEL

In this section, we show how to represent the packages and tourists by a topic model, likes [4], [23], [25] based on Baysian networks, so that the similarity between different packages and tourists can be measured. For better illustration, Table I lists mathematical notations used in this paper.

When designing a travel package, people in travel companies often need to consider the following issues. First, it is necessary to determine the group of target tourists, the travel seasons, and the travel places. Second, one or multiple topics, such as "$Cultural\ travel$" or "$The\ Sunshine\ Trip$", will be chosen for the travel package based on the category of target tourists and the scheduled travel seasons for this package. Each package and landscape can be viewed as a mixture of a number of topics. Then, the landscapes will be determined according to the package topics and the geographic locations of landscapes. Finally, some additional information, such as the information about price, transportation, and accommodations, should be included. According to these main processes and factors, we can formalize the package generation as a "$What-Who-When-Where$"($4W$) problem. Here, each W stands for the package topics, the target tourists, the package seasons and the corresponding landscape located areas, respectively. These four factors are strongly correlated with each other.

Formally, we reprocess the generation of a package in a topic model style. First, we treat the package generation mainly as a landscape drawing problem. These landscapes for the package are drawn from the landscape set one by one, and the package generation is completed after all the landscapes have been chosen. In order to choose 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 the chosen travel area. We call our model for package representation as the TAST (Tourist-Area-Season Topic) model. We should note that, the topic mentioned in TAST is different from the 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. Since these two type of topics can be easily distinguished from the context, we use the same word topic to stand for both of them in this paper.

While the generation processes in TAST are similar to

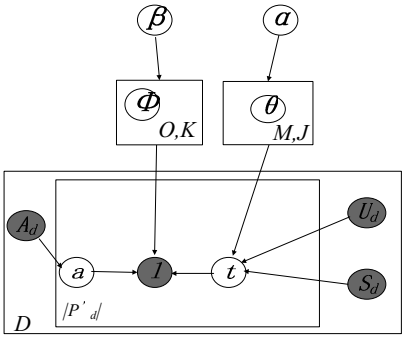


Figure 5. TAST: A graphical model.

those in the text modeling problems for both documents [4], [25] and emails [23], the TAST model is quite different from these traditional ones. 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 auto-correlations among landscapes. The benefit is that the TAST model can describe the package and the tourist interests more precisely, because the nearby landscapes or the landscapes preferred by the same group of tourists tend to have the same topic. In addition, the text modeling has the assumption that the words in a document/email 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.

Mathematically, the generative process corresponds to the hierarchical Bayesian model for TAST is shown in Figure 5, where shaded and unshaded variables indicate observed and latent variables respectively.In the TAST model, the notation $P_d^{'}$ is different from P_d, where P_d is a package in the package set while $P_d^{'}$ is the package for one log in the travel log set. The package log can be distinguished by a vector of three attributes $\langle package\ ID, tourist\ ID, travel\ time\rangle$. This means there is one and only one P_i which is the same as $P_d^{'}$, but there are more than one $P_d^{'}$ that is the same as P_i according to different ($tourist\ ID$, $travel\ time$) pairs. In Figure 5, each package $P_d^{'}$ is represented as a vector of $|P_d^{'}|$ landscapes where each landscape l is chosen from one area a and $a \in A_d$. The set A_d includes the located area(s) for package $P_d^{'}$ and (U_d,S_d) is the specific tourist-season pair. t is a travel topic which is chosen from the topic set T with K 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 each entry in θ and ϕ, can be extracted after inferring this TAST model and estimating the parameters. This variable inference is to "invert" the generative process and "generate" latent variables from given observations. The general idea is to find a latent variable (i.e., topic) setting so as to get a marginal distribution of the travel log set $P^{'}$, which can be computed as:

$$P(P^{'}|\alpha,\beta,U,S,A) = \iint \prod_{i=1}^{M}\prod_{j=1}^{J} P(\theta_{ij}|\alpha) \prod_{i=1}^{O}\prod_{j=1}^{K} P(\phi_{ij}|\beta) \prod_{d=1}^{D}\prod_{i=1}^{|P_d^{'}|}\sum_{t_{di}=1}^{K} (P(t_{di}|\theta_{U_d S_d}) \sum_{a_{di}\in A_d} (P(a_{di}|A_d)P(l_{di}|\phi_{a_{di}t_{di}})))d\phi d\theta$$

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 [16], 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 traveling 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 [6]:

$$\hat{\theta_{ijt}} = \frac{\alpha_t + n_{ijt}}{\Sigma_{k=1}^{K}(\alpha_k + n_{ijk})}, \quad \hat{\phi_{ijl}} = \frac{\beta_l + m_{ijl}}{\Sigma_{k=1}^{|A_i|}(\beta_k + m_{ijk})}$$

where $|A_i|$ is the number of landscapes in area A_i, n_{ijk} is the number of landscape tokens assigned to topic T_k and tourist-season pair (U_i,S_j), and m_{ijk} is the number of tokens of landscape L_k assigned to area-topic pair (A_i,T_j). To better understanding the Gibbs sampling process, let's take the topic assignment for "*Central Park*" as an example, in each iteration, the topic assignment of one "*Central Park*" token depends on not only the topics of the landscapes that traveled by the tourist in the given season but also the topics of the other landscapes located nearby. Besides θ and ϕ, many other posterior probabilities can be derived from Gibbs sampling at the same time, for example, the topic distribution of tourist U_i and package P_i can be estimated by:

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

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

Please note that after the Gibbs sampling, all the tourists and packages can be represented as different topic distribution vectors. By computing the similarity of their topic distribution vectors, we can find the similarity between the corresponding tourists and packages. In addition to this, there are many other benefits of the TAST model, for example, we can learn the popular topics in each season and we can find the popular landscapes for each travel topic or for each topic-area pair.

[6]The detailed inference is omitted due to the space limit.

Table II
AREA SEGMENTATION RESULT.

Area	Provinces/Countries	Landscape Numbers
SC	Guangdong,Guangxi,Macau,Yunnan, Hong Kong,Fujian,Hainan	509
CC	Jiangxi,Guizhou,Sichuan,Hunan,Zhejiang, Jiangsu,Shanghai,Chongqing,Hubei,Anhui	149
NC	Shananxi,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

Another important issue for TAST model is the coverage of each area A_i and each season S_i. There are two extremes: 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, for the first extreme which is too coarse, we can not capture the spatial-temporal auto-correlations. For the second extreme, which is too fine, we will face the overfitting issue and this will makes the Gibbs sampling difficult to converge. To this end, we divide the entire location space in our data set into 7 big areas 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. The detailed area segmentation result is shown in Table II. To make more reasonable season splitting, we assume that most packages are seasonal, and we use an information gain based method [10] to get the season splits such that the travel packages have a relatively stable distribution in each slot. The information entropy of the season S^P is given by $Ent(S^P) = -\sum_{i=1}^{|S^P|} p_i log(p_i)$, 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 in a recursive binary way. In each iteration, we use the weighted average entropy (WAE) to find the best split:

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

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

IV. A COCKTAIL APPROACH ON TRAVEL PACKAGE RECOMMENDATION

In this section, we propose a cocktail approach on personalized travel package recommendation based on the TAST model, which follows a hybrid recommendation strategy 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, the 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.

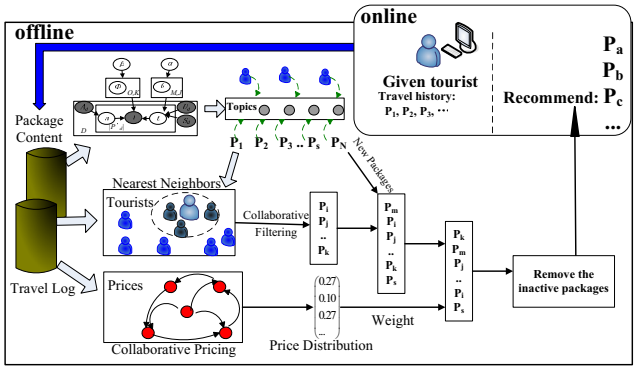


Figure 6. The framework of the cocktail approach on travel package recommendation.

Figure 6 illustrates the framework of the proposed cocktail approach, and each step of this approach is introduced in the following subsections. 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 travel topics of each landscape evolves very slowly, we can update the inference process periodically offline in real-world applications.

A. Seasonal Collaborative Filtering for Tourists

In this subsection, 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 him/her. However, as 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 and they are represented in vectors with the same length. Multiple methods can be used to compute these similarities, such as matrix factorization [21], [20] and graphical distances [9], [11]. Alternatively, a simple but effective way is to use

the *Correlation coefficient* [24], and the similarity between tourist U_m and U_n in season S_j can be computed by:

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

where $\bar{\theta}_{mj}$ is the average topic probability for the tourist-season pair (U_m, S_j)[7], For a given tourist, we can find his/her nearest neighbors by ranking their similarity values. Thus, the packages, which are 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.

B. New Package Problem

In traditional recommender systems, there is a cold-start problem. In other words, it is difficult to recommend new items. For the travel data, as we have explored in Section II, travel packages often have a life cycle and new packages are usually created every year. At the same time, 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's take the year of 2010 as an example. There are 65 new packages in our 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:

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

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. At last, since there is no effective method to learn the topic distributions 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. However, there are few such packages.

[7]If the given tourist U_m has never traveled in season S_j, then his/her total topic distribution ϑ_m^U is used as an alternative throughout this paper.

C. Collaborative Pricing for Optimal Package Recommendations

In this subsection, 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. In addition, there are also time constraints in travel packages, which can be 1-day travel or n-day travel. However, after the analysis in Section II, we find that there is some correlation between the time constraints and the price constraints. Therefore, we focus on studying the impact of the price factor. 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. The method is similar to the one used in [30] for clustering the travel times of taxi drivers. We first sort the prices of the travel logs, and then partition the sorted list PL into several sub-lists 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:

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

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

In the second phase, we mark each price segment as a price state and compute the transition probabilities between them. All the transition probabilities compose a state transition matrix. From the current price state of a given tourist, we predict the next possible price state by the one-step Markov forecasting model. At last, we get the probability distribution on each state, and we use these probabilities as weight to multiply the probabilities of the candidate packages in the rough recommendation list so as to reorder them. After removing the packages that are no longer alive, we get the final recommendation list for the given tourist.

V. Experimental Results

In this section, we evaluate the performances of the cocktail recommendation approach on real-world travel data. Specifically, we demonstrate: (1) the results of the season splitting and price segmentation, (2) the predictive power of the TAST model measured by the perplexity value, (3) the understanding of the topics extracted by the TAST model, (4) a recommendation performance comparison between Cocktail and benchmark methods.

A. 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 part of the test set, and the remaining records were used for training. In all, there are $5,211$ tourists and $22,201$ travel records for 843 packages (1054 landscapes) in the training set, and $1,150$ tourists and 908 packages (1065 landscapes) for testing. There are 65 new packages traveled by 269 tourists in test set. However, only two of these packages are composed completely by new landscapes, and there are 11 new landscapes.

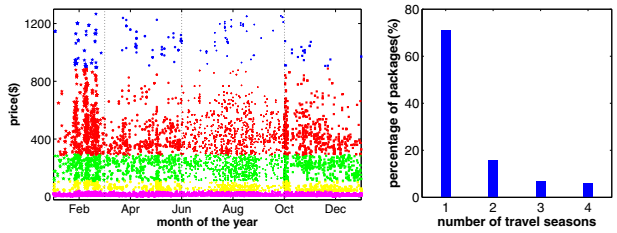
Benchmark Methods. To demonstrate the effectiveness of Cocktail, we compare it with many other methods for both the fitness of the TAST model and the recommendation accuracies. For the fitness purpose, we compare TAST with three related models TAT (Tourist-Area Topic model), TST (Tourist-Season Topic model) and TT (Tourist Topic model), which do not take the season, area, and both season and area factors into consideration, respectively.

For the recommendation accuracies, we compare Cocktail with two other topic model based approaches TTER (a similar cocktail method but based on the TT model) and the TASTcontent (the content based cocktail method where the content similarity between packages and tourists are used instead of using collaborative filtering). For the memory based collaborative filtering, we implemented the user based collaborative filtering method (UCF) [24]. For the model based collaborative filtering, we chose SVD [26]. Since these two methods (i.e., UCF, SVD) only use package level information, to make a more fair comparison, we implemented two similar methods based on landscapes (i.e., LUCF, LSVD). Also, we compared with one graph-based algorithm, ItemRank [15], where a landscape correlation graph is constructed, and for each tourist, the packages are ranked by the expected average steady-state probabilities on their landscapes. Thus, we name this method LItemRank. All the above seven methods (i.e., UCF, SVD, LUCF, LSVD, LItemRank, TTER, TASTcontent) are the benchmarks.

In the following, we choose the fixed Dirichlet distributions with β=0.1 and $\alpha = 50/K$ for topic models, and these settings are widely used in the existing works [16], [23].

B. Evaluation Metrics for Recommendation

In the travel data, there are no explicit ratings for validation. Therefore, we refer to the ranking accuracies. In the experiments, we adopt Degree of Agreement (DOA), Top-K, and the Normalized Discounted Cumulative Gain (NDCG) [18], [22] as the evaluation metrics. All of them are commonly used for ranking accuracies, and they try to characterize the recommendation results from different perspectives: DOA describes the average rank accuracy for the test packages [11]; Top-K indicates the effectiveness of the recommendation from a cumulative way [21]; NDCG evaluates



(a) (b)

Figure 7. The results of season splitting and price segmentation.

the quality of a ranking result in information retrieval by assigning graded content relevance judgments [28].

Since DOA and Top-K used in this paper are similar to [11] and [21] respectively, we only introduce the definition of NDCG. The NDCG metric is evaluated over some k of the top packages on the ranked package list, based on the assumption that highly relevant packages should appear earlier in the recommendation list (have higher ranks) and highly relevant packages are more useful than marginally relevant packages. The NDCG value at the k-th position of the ranking result for a given tourist is computed by:

$$NDCG@k = \frac{RL@k}{IRL@k}, \quad RL@k = (R(P_t, P_1) + \sum_{i=2}^{k} \frac{R(P_t, P_i)}{log_2(i)})$$

where P_t is the test package for the given tourist and IRL@k is the RL@k of an ideal ordering result. The content relevance of two packages P_t and P_i (Here, P_i stands for the i-th package in the recommendation list.) for P_t is $R(P_t, P_i)$ and is defined as $\frac{Num(P_t,P_i)}{|P_t|}$, where $Num(P_t, P_i)$ is the number of co-landscapes for them, and the price is also treated as a landscape. The value of NDCG ranges from 0 to 1 and a higher value indicates better ranking result.

C. Season Splitting and Price Segmentation

In this subsection, we present the results of season splitting and price segmentation as shown in Figure 7. For better illustration, in Figure 7(a), we only show the travel logs with prices lower than $\$1,300$. In the figure, different price segments are represented with different colors and seasons are split by the dashed lines among months. In total, we have 4 seasons including spring, summer, fall, and winter, and 5 price segments. Since almost all the tourists in the data are from South China, this season splitting scheme has well captured the climatic features there.

In Figure 7(a), 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).

What's more, Figure 7(b) describes the relationship between the percentage of the travel packages and the number of scheduled travel seasons. In Figure 7(b), we can see that most of the packages are only traveled in one season during a year, and less than 6% packages are scheduled in the entire year. At last, we should 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

Table III
AN ILLUSTRATION OF SEVERAL TOPICS WITH DIFFERENT SPATIAL-TEMPORAL CHARACTERISTICS.

Topic 4			Topic 14		
Landscape,City	Area	Travel Seasons	Landscape,City	Area	Travel Seasons
Sunflower Garden,Panyu	SC	Spring	Jockey Club,Macau	SC	Entire Year
Long Dear Farm,Shunde	SC	Spring,Summer	Taipa House Museum,Macau	SC	Entire Year
Tenglong Cave Drifting,Jiangmen	SC	Summer	Portuguese Food,Macau	SC	Entire Year
Shunfengshan Park,Shunde	SC	Spring,Summer	Lisboa Casino,Macau	SC	Entire Year
Baomo Gardon,Panyu	SC	Spring	Seaview Bodhisattva,Macau	SC	Entire Year
Longquan Double Falls,Jiangmen	SC	Summer	Friendship Bridge,Macau	SC	Entire Year
Topic 47			**Topic 49**		
Landscape,City	Area	Travel Seasons	Landscape,City	Area	Travel Seasons
Small Yangtes Gorges,Qingyuan	SC	Summer	Wildlife Safari Park,Guangzhou	SC	Entire Year
Summer Palace,Beijing	NC	Spring-Fall	Tian'anmen Square,Beijing	NC	Entire Year
Jiangling Scenery,Shangrao	CC	Summer,Fall	Xiangji Bakery,Panyu	SC	Entire Year
The Great Wall,Beijing	NC	Spring-Fall	Xintiandi Street,Shanghai	CC	Entire Year
Gulong Cave Drifting,Qingyuan	SC	Summer	Beihai Park,Beijing	NC	Entire Year
Acient Town,Jiangwan	CC	Summer,Fall	Bird's Nest Shops,Bangkok	SA	Entire Year

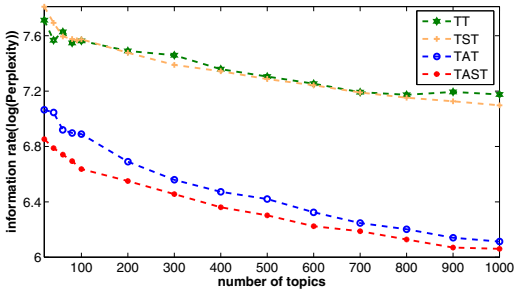


Figure 8. A perplexity comparison.

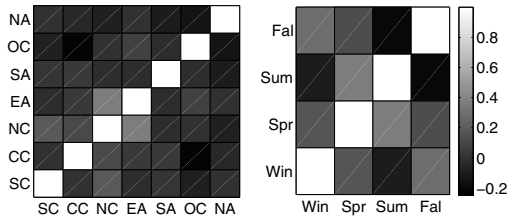


Figure 9. The correlation of topic distributions between different areas (Left)/different seasons(Right). Dark shades indicate lower similarity.

all the packages in the data located in only one of the 7 travel areas. As a result, these statistical results reflect the fact that landscapes in most packages have strong spatial-temporal auto-correlations, and the travel area and travel season segmentation methods are reasonable and effective.

D. Perplexity Comparison

In natural language, the models are often evaluated by perplexity for measuring the goodness of fit. The lower perplexity a model is, the better it predicts the new documents (packages in this paper) [23]. When the tourist U_d, the travel season S_d, and the located areas A_d are given, the perplexity of a previous unseen package log $P_d^{'}$ including landscapes $L_{P_d^{'}}$ can be defined as follows:

$$Perplexity(L_{P_d^{'}}) = exp(-\frac{logP(L_{P_d^{'}}|U_d, S_d, A_d)}{|P_d^{'}|})$$

where $|P_d^{'}|$ is the number of landscapes. In the experiments, four Markov chains were run with different initializations, and the samples at the 1001th iteration were used to estimate θ and ϕ. Here, we report the average information rate (logarithm of perplexity) with different numbers of topics on the data set in Figure 8. As shown in the figure, TAST has significantly better predictive power than three other models, and TT performs the worst since it does not take the spatial or the temporal information into the consideration. While TAT has considered the spatial factor and TST has considered the temporal factor, TAT performs much better than TST. This may be due to the fact that the spatial auto-correlation of landscapes in packages is more obvious than the temporal auto-correlation. Finally, note that the implementation of TT is very similar to the AT model [25]. In order to keep the consistency and to distinguish two different application domains, we use the word TT in this paper to stand for the Tourist Topic model.

E. Travel Topics from the TAST Model

For understanding the travel topics, latent factors inferred by TAST, we mainly focus on studying the relations between the topics and their spatial-temporal characteristics.

Table III shows the highest probability landscapes from four topics in the TAST model trained with 50 topics. For each landscape, we also present its located area and travel seasons. We can see that these four topics stand for the four types of spatial-temporal correlations. First, the prominent landscapes in topic 4 all locate in South China and they are only available in the middle of the year, which means topic 4 has both spatial and temporal correlations. Also, while the landscapes in topic 14 are from different travel packages, all of them are the attractions in the city of Macau, which means this topic also has very strong spatial correlation. However, these landscapes can be traveled throughout the year. In other words, topic 14 only has spatial correlation but temporal correlation. In addition, landscapes in topic 47 and 49 locate in different areas and both topics have no spatial relation. However, landscapes in topic 47 have temporal correlations while landscapes in topic 49 do not have.

Based on spatial-temporal correlations, all the topics can be classified into four types. The existence of all these four

Table IV
A PERFORMANCE COMPARISON: DOA(%).

Alg.	UCF	SVD	LUCF	LSVD	LItemRank	TTER	TASTcontent	Cocktail
DOA(%)	69.96	64.30	88.44	86.85	84.76	89.82	80.00	**92.56**

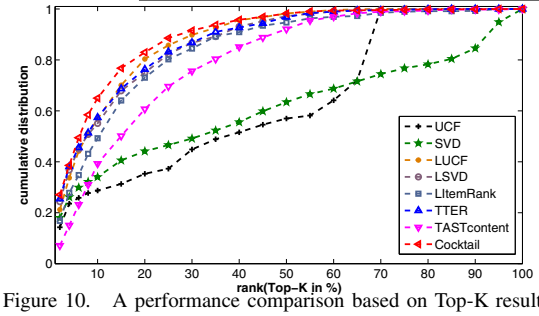


Figure 10. A performance comparison based on Top-K results.

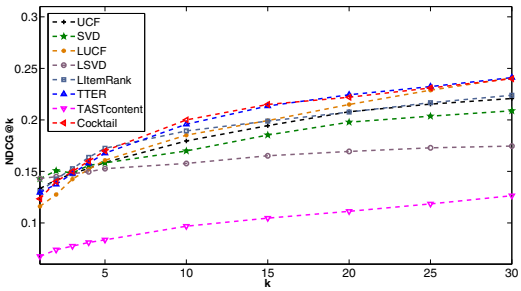


Figure 11. NDCG@k results for different methods.

type topics reveals that TAST model can capture the spatial-temporal correlations existing in the travel data, and these landscapes that are close to each other or with similar travel seasons can be discovered. Meanwhile, the TAST model retains the good quality of the previous topic models for capturing the relations between landscapes locating in different areas and having no special travel season preference.

Furthermore, we show the Pearson Correlations of the topic distributions for different areas/seasons in Figure 9. As shown in Figure 9, different areas/seasons are assigned with different topic distributions. In the left matrix, for most area pairs, there are no obvious travel topic correlations, except for East Asia (EA) and North China (NC). The different types of topic relations 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 travel topics.

F. Recommendation Performances

In this subsection, we present the performance comparison on recommendation accuracies between Cocktail and the benchmark methods. For comparison, we fix topic=100 for TTER, TASTcontent, and Cocktail because the variances of perplexity become less obvious since then, as shown in Figure 8. We also set the number of dimensions as 100 for SVD/LSVD, and set the nearest neighbor size of UCF/LUCF as 1000. For other methods, the neighbors that have a similarity value bigger than 0 are considered. Following [15], the decay factor in LItemRank is also set to be 0.85.

DOA. The average ranking performance of each method is shown in Table IV, where we can see that Cocktail outperforms the benchmark methods. Also, the methods that consider landscape information (i.e., LUCF, LSVD, LItemRank, TTER, TASTcontent, Cocktail) perform much better than those can not (i.e., UCF, SVD). As we have mentioned previously, the reason is that, in this scenario it is harder to find the credible nearest neighbor tourists (and latent interests) only based on the co-traveling packages.

Top-K. In addition, the cumulative distribution of Top-K ranking performances of each method is plotted in Figure 10. As shown in Figure 10, Cocktail still outperforms other methods and the Top-K result is very similar to the DOA result. We should note that there exists a leap in the line for UCF in Figure 10, this is because there are many packages which are not covered by the UCF method and they are given the same default rank.

NDCG. At last, we consider the NDCG scores for different algorithms as shown in Figure 11 with the position k ranging over 1 to 30. Different from DOA/Top-K, in this metric, Cocktail doesn't perform best at the first two positions. However, in all, Cocktail and TTER behave similarly and both of them perform better than other algorithms. Another interesting observation is that, the content based TASTcontent method performs worst. This means the interests of the tourists often change, and it may not be appropriate to only deal with their past preferences.

In all, Cocktail performs better than other methods for all metrics and TTER has the second best performance. Due to the unique characteristics of the travel data, the traditional collaborative filtering methods (UCF and SVD) do not perform well, and meanwhile, they cannot recommend new packages. In general, the methods that consider more information tend to get better performance.

VI. RELATED WORK

To the best of our knowledge, there are few existing works on personalized travel package recommendation. However, there are some recommendation studies in the tourism domain [5], [13], [17], [19], [27], [28], [29].

The related work can be grouped into two categories. In the first category, people target on providing more travel information to help tourists. For instance, Jing et al. developed the VirtualTour system, which is an online travel service to help tourists get more travel information [19]. Also, Wu et al. designed a system using the multimedia technology to generate personalized tourism summary in the form of text, image, video, and news [27].

In the second category, people are focused on the development of recommender systems for tourists. For instance, Carolis et al. [5] developed a mobile recommender system which uses a map for outlining the location and the information of landscapes in a town area. Considering

the travel cost (i.e., the financial cost and the time), Ge et al. [13] provided a focused study of cost-aware tour recommendation. Also, Hao et al. [17] proposed a Location-Topic model by learning the local and global topics to mine the location-representative knowledge from a large collection of travelogues, and used this model to recommend the travel destinations. Moreover, Yin et al. [29] proposed an automatic trip planning framework by leveraging geo-tagged photos and textual travelogues. Finally, Xie et al. [28] proposed a method of composite recommendation of points of interest for tourists according to the tourist's budget.

In all, the above studies mainly focus on mining information from the tourist generated content. Only few of them have considered the relationships of many landscapes [29] and recommend points of interest for tourists [28]. However, these two studies [28], [29] target at generating travel routes for a tourist according to his/her input queries or constraints rather than recommending the existing travel packages.

VII. Concluding Remarks

In this paper, we provided a study of exploiting online travel information for personalized travel package recommendation. Specifically, we first analyzed the unique characteristics of travel packages and developed the Tourist-Area-Season Topic (TAST) model, a Baysian 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 which are inherent in personalized travel package recommendation. Finally, an empirical study was conducted on real-world travel package data. The experimental results demonstrate that the TAST model can capture the unique characteristics of the travel packages, and the cocktail approach can lead to better performances of travel package recommendation.

VIII. Acknowledgements

This research was supported in part by National Natural Science Foundation of China (Grant No.s 61073110 and 71028002), the Key Program of National Natural Science Foundation of China (Grant No. 60933013), the research fund for the Doctoral Program of Higher Education of China (Grant No. 20093402110017), the National Major Special Science & Technology Project (Grant No. 2011ZX04016-071), Fund of Ministry of Education of China (No. 10YJC630065), and National Science Foundation (NSF) via grant numbers CCF-1018151 and IIP-1069258.

References

[1] G. D. Abowd, C. G. Atkeson, J. Hong, and et al. Cyber-guide: A mobile context-aware tour guide. *Wireless Networks*, 3(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 TKDE*, 17(6), pp. 734–749, 2005.

[3] O. Averjanova, F. Ricci, and Q. N. Nguyen. Map-based interaction with a conversational mobile recommender system.In *UBICOMM'08*, pp. 212–218, 2008.

[4] D.M. Blei, Y. N. Andrew, and I.J. Michael. Latent Dirichlet Allocation. *JMLR,3*, pp. 993–1022, 2003.

[5] B. D. Carolis, N. Novielli, V. L. Plantamura, E. Gentile. Generating Comparative descriptions of places of interest in the tourism domain. In *ACM RecSys'09*, pp. 277–280, 2009.

[6] F. Cena, L. Console, and et al. Integrating heterogeneous adaptation techniques to build a flexible and usable mobile tourist guide. *AI Communications*, 19(4), pp. 369–384, 2006.

[7] K. Cheverst, N. Davies, and et al. Developing a context-aware electronic tourist guide: some issues and experiences. In*ACM SIGCHI*, pp. 17–24, 2000.

[8] N. A. C. Cressie. Statistics for spatial data. *Wiley and Sons, ISBN:0471843369*, 1991.

[9] C. Ding, R. Jin, T. Li and H.D. Simon. A learning framework using Green's function and kernel regularization with application to recommender system. In *ACM SIGKDD'07*, pp. 260–269, 2007.

[10] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In IJCAI, pp. 1022–1027, 1993.

[11] F. Fouss, A. Pirotte, J.-M. Renders, etc. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE TKDE*, 19(3), pp. 355–369, 2007.

[12] Y. Ge, H. Xiong, A. Tuzhilin, and Q. Liu. Collaborative filtering with collective training. In *ACM RecSys'11*, 2011.

[13] Y. Ge, Q. Liu, H. Xiong, A. Tuzhilin, and J. Chen. Cost-aware travel tour recommendation. In *ACM SIGKDD'11*, pp. 983–991, 2011.

[14] Y. Ge, H. Xiong, A. Tuzhilin, etc. An energy-efficient mobile recommender system. In *ACM SIGKDD'10*, pp. 899–908, 2010.

[15] M. Gori, and A. Pucci. Itemrank: A random-walk based scoring algorithm for recommender engines. In *IJCAI'07*, pp. 2766–2771, 2007.

[16] T.L. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS'04* vol.101, pp. 5228–5235. 2004.

[17] Q. Hao, R. Cai, C. Wang, etc. Equip tourists with knowledge mined from travelogues. In *WWW'10*, pp. 401–410, 2010.

[18] K. Järvelin, J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4), pp. 422–446, 2002.

[19] F. Jing, L. Zhang, and W. Ma. VirtualTour: an online travel assistant based on high quality images. In *ACM MM'06*, pp. 599–602, 2006.

[20] Y. Koren, R. Bell and C. Volinsky. Matrix Factorization Techniques for Recommender Systems. In *IEEE Computer,* vol.42(8), pp. 30–37, 2009.

[21] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *ACM SIGKDD'08*, pp. 426–434, 2008.

[22] N.N. Liu, Q. Yang. EigenRank: a ranking-oriented approach to collaborative filtering. In *ACM SIGIR'08*, pp. 83–90, 2008.

[23] A. Mccallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *JAIR 30*,pp. 249–272, 2007.

[24] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl. GroupLens: an open architecture for collaborative filtering of netnews. In *ACM CSCW'94*, pp. 175–186, 1994.

[25] M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth. The author-topic model for authors and documents. In *UAI'04*, pp. 487–494, 2004.

[26] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Application of dimensionality reduction in recommender systems-a case study. In *ACM WebKDD Workshop*, pp. 82–90, 2000.

[27] X. Wu, J. Li, S. Neo. Personalized multimedia web summarizer for tourist. In *WWW'08*, pp. 1025–1026, 2008.

[28] M. Xie,L. V. S. Lakshmanan, P. T. Wood. Breaking out of the box of recommendations: from items to packages. In *ACM RecSys'10*, pp. 151–158, 2010.

[29] H. Yin, X. Lu, C. Wang, N. Yu, L. Zhang. Photo2Trip: An Interactive Trip Planning System Based on Geo-Tagged Photos. In *ACM MM'10*, pp. 1579–1582, 2010.

[30] J. Yuan, Y. Zheng, C. Zhang, etc. T-drive: driving directions based on taxi trajectories. In *ACM SIGSPATIAL GIS'10*, pp. 99–108, 2010.