Traveling Route Generation Algorithm Based On LDA and Collaborative Filtering

Peng Cui, Yuming Wang and Chunmei Li*
Department of Computer Technology and Applications
Qinghai University
Xining, China, 810016
e-mail: li_chm0422@sina.com

***Abstract*—With the rapid development of China's economy and the increase in tourism consumption, the number of people in traveling in domestic tourism has increased rapidly each year, and more travelers choose privately customized travel routes, so reasonable travel route is generated based on the actual users' needs has become a hot research spot in the current industry and academia. However, as far as practical application is concerned, the planning of travel routes is a comprehensive and complex task. Reasonable travel routes include comprehensive features such as reasonable travel cities, travel time, transportation methods, and itinerary arrangements. At present, the traditional method is basically that the customer manager can manually plan the suitable travel route for the user through collecting the user's needs, and then modify and adjust by communicating with the customer. The problem that this brings is that the customer manager needs to compare information such as users' needs, travel price, travel time, travel transportation, and scenic spot arrangements when planning numerous travel routes. Obviously, the traditional methods have significant disadvantages such as low efficiency and long time-consuming. Bring a great burden to the staff and it is incompatible with the development of the current industry.**

In order to solve the above problems, we put the historical travel routes collected as data sets in the paper, and a travel route recommendation and generation algorithm based on LDA and collaborative filtering is designed. Reasonable city recommendation list and playing time are the basis and focus of route planning. The paper is based on the many shortcomings in the traditional travel route planning method, and takes the city's recommendation and time planning as the main focuses on work. In this work, different recommendation algorithms were designed, including a recommendation algorithm based on Latent Dirichlet Allocation (LDA) and collaborative filtering. By analyzing the performance of the recommendation algorithm on the data sets, the recommendation algorithm is improved and optimized. The LDA algorithm based on KDE (Kernel Density Estimation) and classification, the collaborative filtering algorithm based on KDE and classification. The final experimental results show that the optimal city list and travel time generated by the recommended algorithm are more reasonable and satisfy the actual use of the user.

Keywords-LDA Model; Collaborative Filtering; KDE Algorithm; Recommended Algorithm; Machine Learning

I. INTRODUCTION

Tourism industry has become an important part of national economy within the rapid development of China's national economy in these years, and the number of travelers has also been gradually increasing. According to the data shown by National Bureau of Statistics, the consumption brought by tourism has also increased year by year. The tourism industry has shown an accelerated convergence of online and offline. Traditional travel agencies have been unable to meet consumers' need and development of modern tourism. Based on the above situation, the online development mode of tourism has become a research hotspot in academia and industry. At present, the development carrier of online tourism is mainly online travel websites

DOI: 10.21307/ijanmc-2019-021

(such as: tuniu.com, tongcheng.com, etc.). The traditional travel website was designed by B/S mode [1], which can provide consumers with a large amount of tourist information, however, the travel route displayed on the website is manually planned by sales account manager after collecting users' requirements. Therefore, this has created two problems. On the one hand, the traditional method of artificially planning travel routes has low productivity and cannot meet the development of tourism. So how to automatically generate a reasonable travel route with intelligent algorithms according to users' needs has become an urgent problem to be solved. On the other hand, traditional travel routes can no longer meet the individual needs of each user, so how to reduce the blindness and randomness of route arrangement, then provide customers with personalized travel routes, thus providing more travel options for users to choose has gradually become a research hotspot of relevant enterprises and disciplines.

In recent years, algorithms for travel route generation, LDA, and collaborative filtering have been reported many times. Ma Zhangbao et al. [2], who began with the space decision-making of tourism travel, studied methods and techniques of the tourism travel decision support system, and then proposed the operational model of travel combining space and time and the LBS model of tourism travel route. However, the model proposed in their paper focuses on querying attractions and hotels based on space, time and location service, and then provide users with query service of the optimal destination as well as the scheme to arrive at destination. But this way can not generate a complete travel route according to the user's demand; it was not suitable for applying in the actual context. Jin Baohui et al. [3] designed a travel route choice model based on Regret Theory and figured out the deficiency of Expect Utility Theory and Prospect Theory via comparison, and then proposed a simpler travel route choice model. This paper focuses on description of tourists' behavior in selecting routes, and then presents a selection model for tourists under uncertain conditions. It didn't do comprehensive comparison and sorting for travel routes, even didn't discuss about actual problems. Also, travel information provided by online travel sites, which provides big data source, were totally ignored in their work. Chunjing Xiao, et al. has proposed a travel route recommendation method based on dynamic clustering. This method firstly analyzes different characteristics of tourism data and other standard data. Secondly, it uses the variable long time window obtained by dynamic clustering to divide the tourist interaction history. The potential Dirichlet distribution (LDA) is used to extract probability topic distribution of each stage, and the user interest drift model is established by combining the time penalty weights. Finally, the route recommendation is completed according to the candidate topic and probability topic relevance of tourists [4]. Although this method has good recommendation accuracy, it focuses on recommending possible routes for users under the premise of there are some user interest sets and numerous travel routes, at the same time, it does not focus on the study of travel route generation. Wang Hui et al. has proposed the solution of ant colony algorithm in the application of travel route planning, they discussed the application in vehicle routing problem based on ant colony algorithm and completed the travel of 201 5A scenic spots in the country that using the shortest time. However, this paper does not study the planning and generation of travel routes that meet needs according to user's preference conditions [5]. HOU Le [6] et al. has proposed an optimization algorithm based on iterated local search (ILS) and cuckoo search (CS). This algorithm firstly uses ILS to solve tourist attractions and initial travel routes. Then, the CS algorithm is used to minimize time cost of travel route while satisfying both the time window constraint of tourist attraction and the total number of attractions. The main problem solved by the algorithm is a complete route of the shortest travel time required given the tourist attractions. The research focuses on minimizing the time of travel routes; in the meanwhile, it has not completed research from city list generation to the development of route plan.

Research and application of recommendation algorithms such as collaborative filtering and LDA (Latent Dirichlet allocation) are reported at home and abroad. Yajun L [7] and others wrote a review of collaborative filtering recommendation techniques. In this paper，what they have done were summarized and compared the collaborative filtering algorithms. This paper reviews the related research of collaborative filtering. Firstly, it expounds the connotation of collaborative filtering and its main situation, including sparsity, multi-content and scalability, and then detail the solutions for domestic and foreign scholars. This article is very helpful for the study and research of collaborative filtering algorithms. Qiang C, et al. has proposed a recommendation algorithm based on label and collaborative filtering. The label is used as information embodying user interest preferences and resource characteristics. The user and resource tag feature vectors are generated based on the multi-dimensional relationship between users, tags and resources. Finally, based on prediction preferences, sorting values will produces Top-N recommendations. Then the collaborative filtering algorithm is applied to the recommendation of personalized resources [8]. One of the most successful applications of collaborative filtering algorithms in foreign countries is the Amazon online website. Amazon's G Linden [9] and others proposed an item-based collaborative filtering algorithm, which is well suited for comparing similar items rather than comparing similar users. The number of items is much larger than the actual number of users, resulting in high quality recommendations. Regarding the LDA algorithm, DM Blei et al. [10] proposed a three-level hierarchical Bayesian model in 2003, and proposed an efficient approximate reasoning technique based on variational method and an EM algorithm for empirical Bayesian parameter estimation. The LDA algorithm was successfully applied to the fields of text modeling, text categorization, topic extraction, etc., and mixed with unigrams and probabilistic LSI models. R Krestel et al. [11] successfully applied the LDA algorithm to the field of tag recommendation. They used the LDA algorithm to mine the user tags under the same theme, and then recommended the new tags to the user as a search condition, which improved the search efficiency.

From the current research status at home and abroad, there is no relevant scholars and enterprises can provide a feasible and accurate method to meet actual requirements. The current research results focus on the recommended method of designing travel routes under the premise of mastering user information and historical route data, which is, recommending the travel route in historical routes through user's historical information, so for the new user's demand, it can't generate a new route that meets the user's needs. At the same time, through the learning and researching for collaborative filtering and LDA algorithm, it is found that these algorithms are feasible and applied in this paper. According to that, we will show the method of recommendation and generation of tourist routes based on LDA and collaborative filtering below.

II. RELATED WORK

The planning of a travel route is a complex and comprehensive process that requires consideration of many factors, such as user’s demand, the price of route, interest arrangement, and transportation. The basic theory of route planning and generation involves multiple disciplines, including data mining, statistical machine learning, network search, pattern recognition, and spatial data mining. A scientific travel route can display as many tourist attractions and landscapes as possible to visitors, thereby improving satisfaction and happiness of tourists and promoting the long-term development of tourism industry. In recent years, with the rapid development of artificial intelligence technology, route planning algorithms such as genetic algorithm, particle swarm optimization algorithm, simulated annealing algorithm, ant colony algorithm and immune algorithm have been emerged. The planning and generation of a travel route mainly involves generating recommended city according to user's needs, and reasonably planning the playing time of recommended city.

This paper takes the collected historical travel route datasets of Japan as researching object, mainly studies the recommendation and generation scheme of travel city time-space list in the travel route planning. It is proposed to use LDA and collaborative filtering to design the travel city recommendation algorithm, using KDE algorithm to optimize the playing time of each city, and then generate a time-space list of user's playing city. In the experimental part, the results of topic city model based LDA and different travel route recommendation algorithms are introduced in detail. The relevant city error rate of topic city model based LDA under different parameters is compared and the optimal model parameters are obtained. Finally, the performance of different recommendation algorithms is evaluated and analyzed.

III. TRAVEL ROUTE RECOMMENDATION AND GENERATION SYSTEM

The travel route recommendation algorithm based on KDE and classification mainly includes three modules. They are data preprocessing and feature extraction module, playing time estimation module based on KDE, topic city generation module based on LDA and travel route generation module or recommended city generation module based on collaborative filtering. The data preprocessing and feature extraction module mainly transforms the original data set into a travel route text set through operations such as data cleaning, classification and feature extraction, that is, it conforms to the input format of LDA model, such as the document-content distribution format. The original data set comes from the travel historical data set of Japan, and there are about 5,000 travel routes. The playing time estimation module based on KDE mainly uses the KDE algorithm to calculate users' total playing time and the playing time of input cities, improving the accuracy of the playing time and the quality of recommended algorithm. The topic city generation module based on LDA is the core module of entire algorithm. In this module, the topic-probability distribution under the travel route text and the characteristic city probability distribution under each topic are calculated through established travel city topic model based on LDA. In turn, the probability distribution of characteristic cities is converted into a list of recommended cities. The topic city generation module based on collaborative filtering is also the core module of entire algorithm. In this module, the list of recommended cities satisfying conditions is calculated through the collaborative filtering algorithm. The travel route generation module is the total output module of algorithm. After processing the output result of previous module, a complete travel route is finally formed, including users' total playing time, the list of travel cities, and the list of playing time of each city. The system structure of algorithm is shown in Figure 1 and Figure 2.

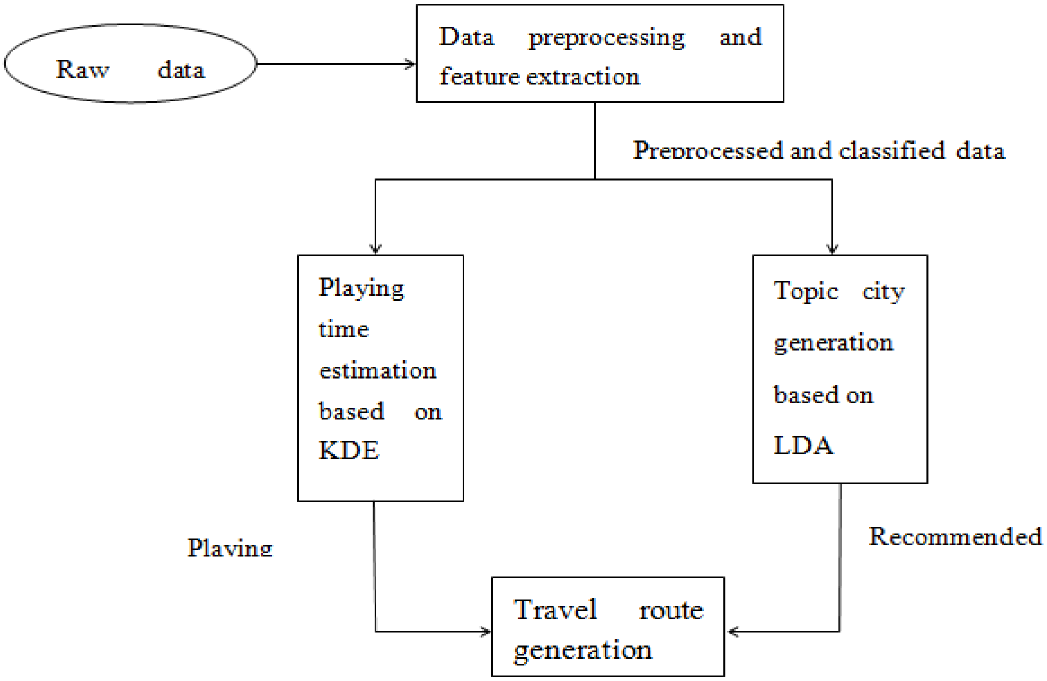
Raw data
Data preprocessing and feature extraction
Preprocessed and classified data
Playing time estimation based on KDE
Topic city generation based on LDA
Playing
Recommended
Travel route generation

Figure 1. LDA travel route recommendation algorithm based on KDE and classification

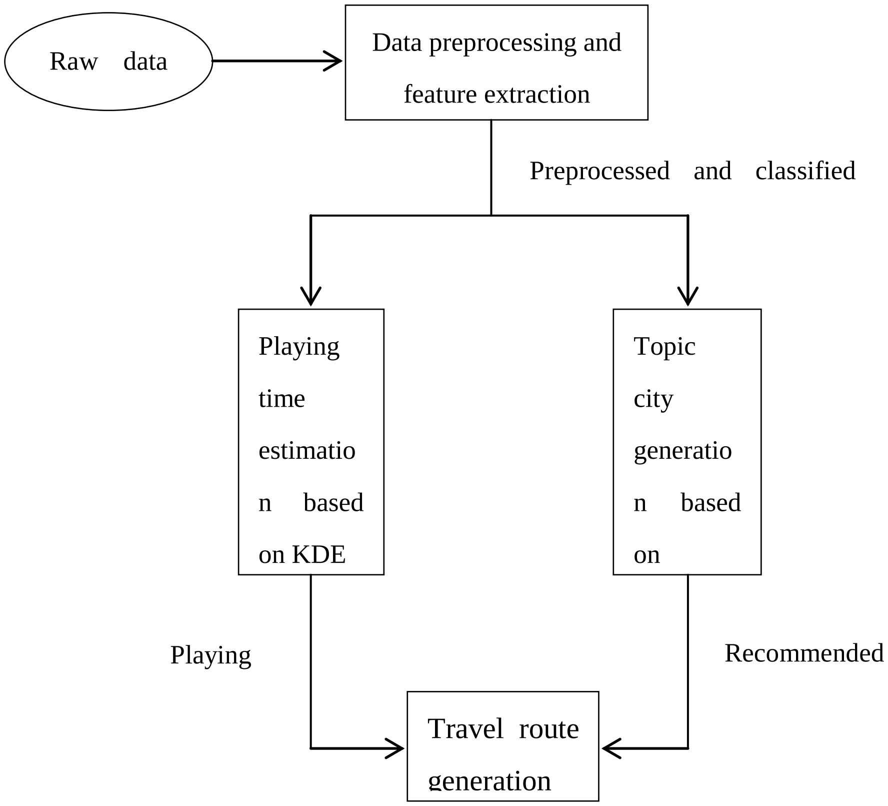
Raw data
Data preprocessing and feature extraction
Preprocessed and classified
Playing time estimation based on KDE
Topic city generation based on
Playing
Recommended
Travel route generation

Figure 2. Collaborative filtering travel route recommendation algorithm based on KDE and Classification

A. Data Preprocessing and Feature Extraction Module

Data preprocessing is basis for algorithm to get good training and output results. In the data cleaning and preprocessing module, feature extraction and data classification are mainly completed. The given original data set is mainly json format travel route data. Each route is an ordered list of multiple city lists. The attributes of each city list include the city ID (id), trip name or city name (name), type, travel time (travel_times or transit_time). Data cleaning is used to extract useful feature data in the data set and complete the missing data. Then the extracted data is sorted according to specific rules, where we classify according to the number of cities of route. Finally, through data preprocessing, the data is organized into a data set that can be used as the LDA model input, such as a document-content distribution format. The specific data preprocessing steps are:

1) Reading the json data file using python code, the city name (name), travel time (travel_times), and route number (plan_id) of each route are read;

2) Calculating the number of writing for each city, the specific number of writing = the total playing time (hour) / 4;

3) Writing the extracted features into different output files according to a specific format ([number of lines, route id, city name]);

4) According to the number of cities in each route, the output will be classified according to the number of cities 4,4-5, 6-7, 8-10, 10, and stored in the corresponding files;

5) The writing of data is completed and the file is saved.

TABLE I. ROUTE BASIC ATTRIBUTE TABLE

	id	**name**	**plan_id**	**type**	**hours**	**daysep**
meaning	City id	City name	Route id	Route type	Playing time	The flag of end of day
Value type	string	string	string	string	list	bool
example	‘263’	‘Osaka’	‘3799’	‘place’	[4.0,8.0]	true

B. Playing time estimation module based on KDE

In general, the total time for users to play is calculated based on people's experience theoretics to formulate specific rules, for example, total hours of playing (days) = total time of playing(days) * playing time of every day (8 hours). The playing time of city that the user wants to go to is calculated by multiplying the probability of topic distribution obtained by LDA by the total playing time. In practical applications, it is found that this method does not have a certain degree of flexibility and cannot adapt to all user inputs. The resulting time error is relatively large, resulting in poor recommendation quality. Therefore, in this paper, we decided to use the KDE (Kernel Density Estimation) algorithm to estimate total time for users to play and the playing time of city that users want to go to, improving the recommendation quality. Assuming that t1 , t2 , … tn are n samples of total playing time t, and the probability density function of total playing time is $f(t)$, the kernel density estimation of $f(t)$ is:

$$\hat{f}_h(t) = \frac{1}{n}\sum_{i=1}^{n} K_n(t-t_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-t_i}{h}\right) \tag{1}$$

Where h is the bandwidth, n is the number of samples, and K () is the kernel function.

The algorithm steps for playing time estimation based on KDE are:

1) According to the number of days of playing, the historical routes will be categorized into five categories: 1-3 days, 4-5 days, 6-7 days, 8-10 days and 10 days or more;

2) Determining the corresponding route data category according to the number of days of playing input by users;

3) Reading the playing time of each route of a specific category, and saving it as a list A;

4) Using list A as the input data of kernel density estimation function to obtain the kernel density estimation function;

5) Randomly sample a function value as the total time for user to play, expressed as H, according to the obtained kernel density function;

6) Repeat the above steps to obtain the list of playing time G of input cities.

According to the above algorithm, we can get total time for user to play, expressed as H and the list of playing time for input cities. These two values will be used later in the topic city generation module based on LDA.

C. Topic city generation module based on LDA

The LDA model is a probabilistic topic model for modeling discrete data sets (such as document sets). LDA is essentially an unsupervised machine learning model that can express high-dimensional text word space as low-dimensional topic space, ignoring text-related category information. The LDA model gets a brief description of document by making topic modeling of document set, retaining the essential statistical information and helping to efficiently process large-scale document sets [12]. In general, before applying LDA model, it is necessary to satisfy the premise that the document is composed of a number of latent topics, which are composed of a number of specific words in the text, ignoring the order of words and syntactic structure in the document. For the travel route dataset of this paper, after data preprocessing and feature extraction, a document set containing discrete city names is formed. Each document is composed of a number of travel cities. There is no syntactic structure in the document, and words are not specific order. And it is in accordance with premise and data requirements of LDA model. In this paper, according to preprocessed travel route datasets, the training set of travel routes is characterized by dimensionality reduction, and the training set is expressed as the form of topic probability, so that a specific topic city is extracted from the topic probability list to form a recommendation city list.

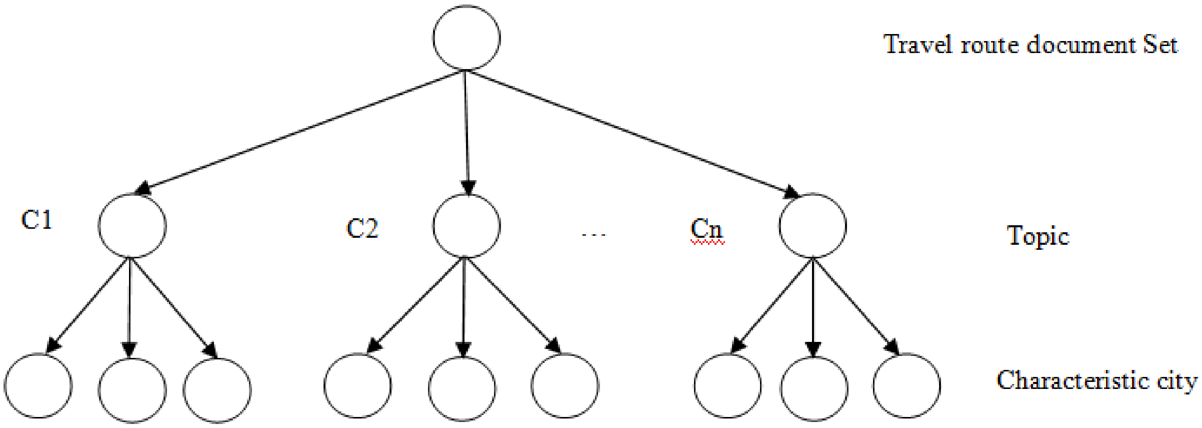

Figure 3. Travel city topic model based on LDA

Figure 3 above shows the established travel city topic model based on LDA. There are three layers in the model, followed by the document collection layer of travel city, topic layer, and characteristic city layer. The process of travel city topic model based on LDA generates a feature city as follows:

1) For topic c, a word polynomial distribution vector φ on the topic is obtained based on Dirichlet distribution Dir (β);

2) The number of words N obtained from the Poisson distribution P;

3) According to the Dirichlet distribution Dir (α), a topic distribution probability vector θ of the text is obtained;

4) For each word w n in the text N words:

a) Polynomial distribution from θ Multinomial (θ) randomly selects a topic z;

b) Select a word as w n from the polynomial conditional probability distribution Multinomial(φ) of topic z.

To obtain the probability distribution of a characteristic city, we need to use model parameter estimation methods to estimate word probability distribution under each topic and topic probability distribution of each text. The more commonly used parameter estimation methods are the expected propagation algorithm, variational Bayesian inference and Gibbs sampling [13] [14]. The high-efficiency Gibbs sampling method is used in this paper estimates the probability distribution of a characteristic city through the Gibbs sampling method in the case of a known travel route data text set. According to LDA model, we can get the probability of a text:

$$p(\omega \mid \alpha, \beta) = \int p(\theta) \mid \alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(\omega_n \mid z_n, \beta)) d\theta \tag{2}$$

Using the Gibbs sampling method, the topic of each word is sampled, and the parameter estimation problem can be converted into calculating the conditional probability of topic sequence under word sequence.

$$p(z_i = k \mid \overrightarrow{z_{\neg i}}, \vec{\omega}) = \frac{p(\vec{\omega}, \vec{z})}{p(\vec{\omega}, \overrightarrow{z_{\neg i}})} \propto \frac{n_{k,\neg,i}^{t} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg,i}^{t} + \beta_t} (n_{k,\neg,i}^{t} + \alpha_k) \tag{3}$$

In the above expression, z_i represents the topic variable corresponding to the i-th word; $\overrightarrow{z_{\neg i}}$ is the i-th word is not included in the expression; n_k^t is the number of occurrences of the word t in the topic k is represented; β_t is the prior probability of Dirichlet of the word t; and n_m^k represents the number of topic k in the text m. α_k is the prior probability of Dirichlet of topic k. Based on the above calculation results and the topic number of each word obtained, parameters to be calculated can be calculated by the following equation:

$$\phi_{k,t} = \frac{n_k^t + \beta_t}{\sum_{t=1}^{V} n_k^t + \beta_t} \tag{4}$$

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k} \tag{5}$$

ϕ_k, t represents the probability of word t in the topic k; $\theta_{m,k}$ represents the probability of topic k in the text m. The input and output of travel city topic model based on LDA is shown in the following table 2:

TABLE II. INPUT AND OUTPUT OF TRAVEL CITY TOPIC MODEL BASED ON LDA

input：preprocessed and classified travel route text set (one route for one line) **The number of topic K, hyperparameters α and β**
output： 1. Topic number assigned to each word of each text 2. Topic probability distribution θ for each text 3. Characteristic city probability distribution for each topic 4. Word id mapping table in the program 5.Top-N feature city words sorted from top to bottom for each topic

D. Travel City Generation Module Based on Collaborative Filtering

Because there are a large number of travel routes in the data set of this research, we can regard each travel route as a user before applying the collaborative filtering recommendation algorithm and consider each travel city in each route as a item. Obviously the number of items in this research is much larger than the number of users, so we use a project-based collaborative filtering recommendation algorithm. The algorithm of a travel city generation module based on collaborative filtering is divided into the following three steps:

1) Establish a route-city scoring dictionary. According to the actual situation, the score here is playing time (hours) of each city. Based on the pre-processed travel route text set, a route-city scoring dictionary was established. The key value of dictionary is the route number, value is also a dictionary, the key is the city name, and value is playing time of city. The format is as follows:

Dic={‘route1’:{‘city-1’: playing time-1,‘city-2’: playtime-2,…,‘city-n’: playtime-n},‘route 2’:{‘city-1’: playtime-1,‘city-2’: playing time-2,…,‘city-n’:playing time-n},…, ‘route-n’:{‘city-1’: playtime-1,‘city-2’: playing time-2,…,‘city-n’: playing time-n}}

2) Calculating the similarity between cities and getting a list of similar cities (neighbors) in each city. In calculating similarity, we use the Euclidean distance to measure the similarity between cities;

3) Generateing a list of recommended cities. A weighted sum of all the cities in the set of city neighborhoods is obtained, and the time for the target route to all cities is finally obtained. After playing time set is sorted, the top-N list is taken as city recommendation list.

E. Travel route generation module

The travel route generation module is an integrated module and an output module of the entire algorithm. Through playing time estimation module based on KDE, the total time for users to play H and the playing time list G for input cities can be obtained.The topic city generation module based on the LDA can be used to get the probabilistic distribution of characteristic city—recommended cities list. We need to normalize the probability of extracted topical city, find out playing time of recommended city based on processed probability value, and finally form a complete travel route. The main process of travel route generation module is as follows:

```
a.  rest ← H – sum(G) #Calculating total playing time of recommended cities list
b.  sum_prop ← 0 #Assigning the total probability value =0
c.  recom_list=get_recom()  #Getting a list of recommended cities，the form: [[city-1，probability value-1],[city2，probability value-2],…]
d.  trip_list ← null   #Assigning the route list to null
e.  for i←0 to size(Recommended city list size)
        do sum_prop←sum_prop+ recom_list [i][1]
        repeat
f.  for i←0 to size(Recommended city list size)
        do recom_list [i][1]←recom_list [i][1]/sum_prop * rest
        repeat
g.  for i←0 to size(Recommended city list size)
        do trip_list [i]←recom_list [i]
        repeat
h.  Add the list of cities entered by the user and their playing time to trip_list
i.  return trip_list
```

Through the travel route generation module, you can get a complete travel route. The specific route format is [[city-1, playing time-1], [city-2, playing time-2]… [city-n, playing time-n]].

IV. THE RESULT AND ANALYSIS OF EXPERIMENT

The evaluation of experimental results is an important work, this chapter mainly shows and evaluates the experimental results of different recommended algorithms, including the results of the topic city generation based on LDA, the results of the LDA travel route recommendation algorithm based on KDE and classification, the results of the collaborative filtering travel route recommendation algorithm based on KDE and classification, The performance of different travel route recommendation and generation algorithms based on LDA and the relevant city error rate are compared under different parameters. In recommendation field, commonly used evaluation indicators include recall rate and precision rate [15][16]. Generally, the accuracy of the recommended algorithm is evaluated by the recall rate and precision rate. In e-commerce systems, the conceptual formulas for recall rate and precision rate are as follows [17]:

Precision rate = the number of items user likes / the number of items recommended by the system;

Recall rate = the number of all user's favorite items in the recommended list / the number of all user's favorite items in the system

Based on the concept and calculation methods of precision rate and recall rate, combined with the research content of this paper, we propose two evaluation indicators of the relevant city error rate and route correlation rate, which are used to evaluate the results of topic city generation model based LDA and route generation results of recommendation algorithm respectively. In popular terms, the relevant city error rate is the probability that a tourist city is classified as a wrong topic (route). Here we use P (e) to represent, which can be calculated by the following formula:

$$p(e)=\frac{c_i}{A_i} \tag{6}$$

In the above formula, Ci represents the number of tourist cities that are classified as the wrong topic in the probability distribution of the i-th topic, that is, in the historical routes, the city is not in the same route as any other city in the topic city. Ai represents the total number of cities in the probability distribution of the i-th topic. Therefore, the lower the relevant city error rate, the higher the quality of the model output, the more easily accepted. In the practical application, the related city error rate is generally not more than 0.2.

According to the relevant city error rate above, we can get the route correlation rate calculation method. Here, we use R (t) to represent:

$$R(t)=1-\frac{t_i}{T_i^2} \tag{7}$$

In the above formula, t_i represents the number of recommended cities that are classified as wrong routes in the i-th generation route, that is, in the historical routes, the city is not in the same route as any other city in the generation route. T_i represents the total number of cities in the i-th generation route. Because in the recommendation process, if there are cities that have no relevance with other cities in the recommended route, it is often unacceptable. Therefore, the higher the route correlation rate, the better the performance of the route recommendation and generation algorithm, the more consistent with the user's expectations. In practical applications, the route correlation rate is generally not less than 80%.

A. *The evaluation of topic city generation model based LDA*

The value of topic K of LDA model, the number of iterations, and the hyperparameters α and β all affect the probability distribution of the final topic city. Therefore, in order to obtain the optimal topic city probability distribution, we examine the effect of probability distribution of the topic city under different parameters. In order to ensure the uniformity of the experimental premises, the sample set of all the experimental results below is a set of 8-10 tourist route texts.

1) *Experimental results under different hyperparameter α*

We set the initial value of topic K = 50, the number of iterations: niter = 500, the hyperparameter β = 0.1, then the hyperparameter α takes 5, 10, 15, until 50. Table 3 and Figure 4 below show the experimental results for different values of hyperparameter α. From the experimental results, it can be seen that the value of the optimal hyperparameter α is 25.

TABLE III. THE EXPERIMENTAL RESULTS OF DIFFERENT VALUES OF HYPERPARAMETER Α

α	5	10	15	20	25	30	35	40	45	50
log	4.72	4.16	4.02	3.38	3.21	3.16	3.82	4.12	4.68	5.16
p(e)	0.282	0.254	0.236	0.192	0.166	0.171	0.179	0.216	0.249	0.288

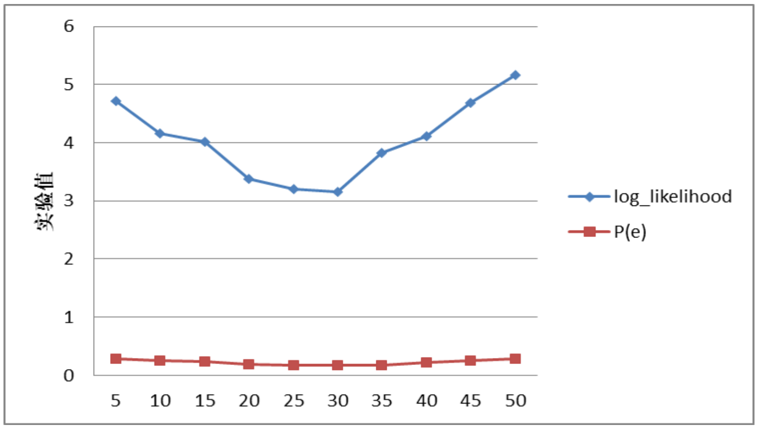

Figure 4. The experimental results of different values of hyperparameter α

2) Experimental results under different hyperparameter β

We set the initial number of topic K = 50, the number of iterations: niter = 500, the hyperparameter α = 25, then the hyperparameter β takes 0.01, 0.05, 0.1, until 0.50. Table 4 and Figure 5 below show the experimental results for different values of hyperparameter β. From the experimental results, it can be seen that the value of the optimal hyperparameter β is 0.15.

TABLE IV. THE EXPERIMENTAL RESULTS OF DIFFERENT VALUES OF HYPERPARAMETER Β

β	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.50
log	5.62	4.42	4.02	3.32	4.23	5.10	5.82	5.92	6.58	7.21
p(e)	0.282	0.254	0.236	0.172	0.198	0.216	0.232	0.299	0.328	0.356

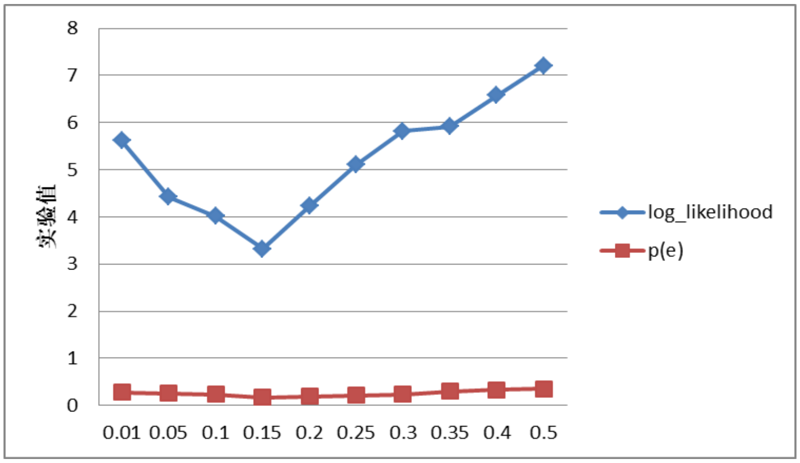

Figure 5. The experimental results of different values of hyperparameter β

3) Experimental results under different number of topic K

We set the number of iterations: niter = 500, the hyperparameter α = 25, β =0.15, then the value of topic K takes 4, 6, 8, until 22. Table 5 and figure 6 below show the experimental results for different number of topic K. From the experimental results, it can be seen that the optimal number of K is 12.

TABLE V. THE EXPERIMENTAL RESULTS OF DIFFERENT NUMBER OF TOPIC K

k	4	6	8	10	12	14	16	18	20	22
log	5.62	4.42	4.02	3.32	3.26	5.10	5.82	5.92	6.58	7.21
p(e)	0.223	0.214	0.205	0.196	0.182	0.226	0.265	0.314	0.408	0.516

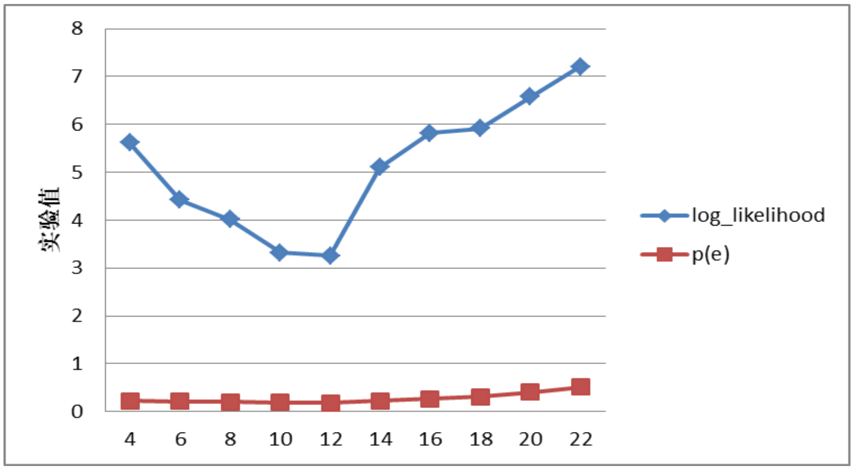

Figure 6. The experimental results of different number of topic K

4) Experimental results under different number of iterations n

We set the initial number of topic K = 12, the hyperparameter α = 25, β =0.15, then the number of iterations take 300,400,500, until 1200. Table 6 and figure 7 below show the experimental results for different number of iterations n. From the experimental results, it can be seen that the optimal number of iterations is 900.

TABLE VI. THE EXPERIMENTAL RESULTS OF DIFFERENT NUMBER OF ITERATIONS N

n	300	400	500	600	700	800	900	1000	1100	1200
log	6.15	5.86	4.02	3.98	3.82	3.64	3.12	3.31	3.41	3.53
p(e)	0.332	0.308	0.275	0.262	0.236	0.214	0.161	0.172	0.181	0.194

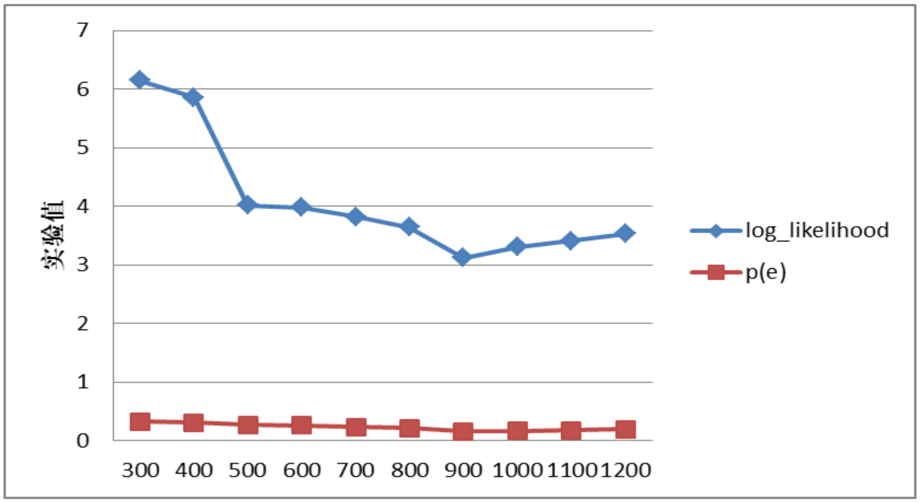

Figure 7. The experimental results of different number of iterations n

From the above experimental results, it can be concluded that the optimal parameters of topic city generation model based LDA are k=12, hyperparameter α=25, β=0.15, and the number of iterations n=900. Under the optimal parameters, the relevant city error rate is 0.161, which is an acceptable error rate in practical applications.

B. Evaluation of LDA travel route recommendation algorithm based on KDE and classification

In order to reduce contingency of experimental results and improve confidence of experimental results, in the experiment evaluation of this section, we carry out the following experimental steps:

1) Randomly generating 50 groups of input city list and playing time;

2) Using the random generated input city list and playing time as input to the recommendation algorithm

3) Recording the output of algorithm obtained from 50 sets of input data, and taking the average value of relevant city error rate of 50 sets of experiments, denoted as Ei

4) Repeating the above steps (1)-(3) for 10 times to obtain the value of Ei for each time.

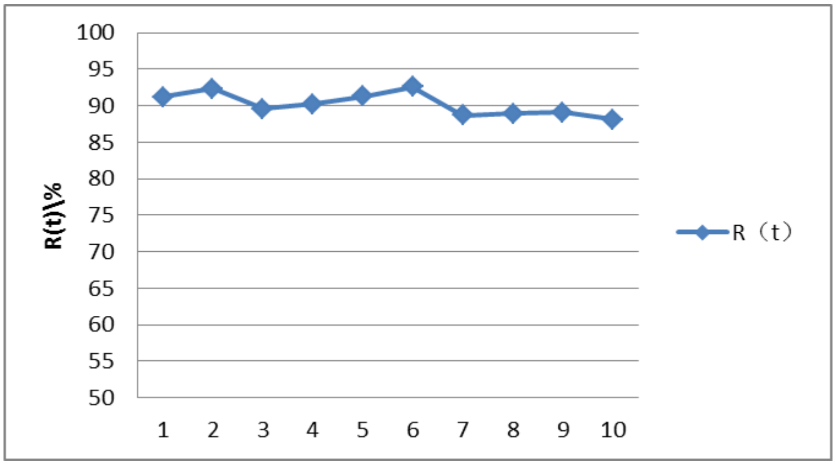

Figure 8. The route correlation rate of LDA travel route recommendation algorithm based on KDE and classification

Figure 8 shows the experimental results obtained in 10 independent experiments. It can be seen from the figure that the value of route correlation rate has been stable at around 90%, because the topic city generation model based LDA has a certain degree of randomness in generating the recommended city list, there will be irrelevant cities in the resulting travel routes. However, by observing experimental results, the route correlation rate of generated travel routes is about 90%, which is within the normal error range. Therefore, the performance of LDA travel route recommendation algorithm based on KDE and classification is relatively good, which conform to practical applications.

C. Evaluation of collaborative filtering travel route recommendation algorithm based on KDE and classification

In order to reduce the contingency of experimental results and improve the confidence of experimental results, in the experimental evaluation of this section, we also carry out following experimental steps:

1) Randomly generating 50 groups of input city list and playing time;

2) Using the randomly generated input city list and playing time as input to the recommendation algorithm

3) Recording the output of algorithm obtained from 50 sets of input data, and taking the average value of relevant city error rate of 50 sets of experiments, denoted as Ei

4) Repeating the above steps (1)-(3) for 10 times to obtain the value of Ei for each time.

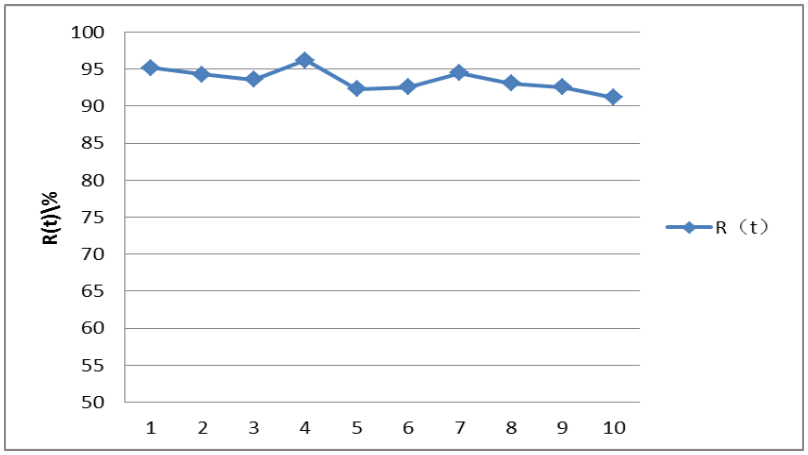

Figure 9. The route correlation rate of collaborative filtering travel route recommendation algorithm based on KDE and classification

Figure 9 shows the experimental results obtained in 10 independent experiments. It can be seen from the figure that the value of the route correlation rate has been stable at around 95%, compared with experimental results of LDA travel route recommendation algorithm based on KDE and classification. The correlation rate obtained by collaborative filtering algorithm is higher, but the difference is not large. This is because the core of collaborative filtering algorithm is based on traditional statistical algorithms, and there is no randomness. The route correlation rate with an error rate of approximately 5% is due to the irrelevance of user s' input city list.. Because in the actual application, cities list input by the user may not exist in the historical routes itself, which is due to systematic error caused by the incompleteness of historical data, which may not be considered in our experiments.

D. Comparison of algorithm output results

TABLE VII. THE OUTPUT RESULTS OF DIFFERENT ALGORITHM

The input of algorithm	total days of travel	7
	cities that user wants to go	Osaka, Nagoya
The output of algorithm	No improved LDA recommended algorithm	[Naoshima: 2.5, Yamanashi: 1.8, Osaka: 56.4, Nagoya: 29.8]
	No improved collaborative filtering recommendation algorithm	[Yakushima: 12.5, Naoshima: 8.6, Osaka: 42.8, Nagoya: 26.2]
	LDA travel route recommendation algorithm based on KDE and classification	[Kyoto: 42.4, Nakafurano-cho: 3.9, Osaka: 15.5, Nagoya: 16.5]
	collaborative filtering travel route recommendation algorithm based on KDE and classification	[Kyoto: 24.2, Tokyo: 20.3, Osaka: 15.5, Nagoya: 16.5]

Table 7 above is results of travel route generation under different recommendation algorithms. It can be seen from experimental results that the improved travel route recommendation algorithm is significantly better than the no-improved algorithm in the playing time schedule. The playing time is more reasonable than previous algorithm. The situation that playing time is too short has reduced, reaching our expected goal.

E. Summary of experimental results

In order to evaluate model results and recommended algorithms, this chapter first proposed new evaluation indicator based on recall rate and precision rate realization principles, relevant city error rate and route correlation rate. Then, the influences of different number of topic K, number of iterations, the hyperparameters α and β on the LDA topic city generation model are introduced. After many experiments, the optimal model parameters are determined to be k=12, α=25, β = 0.15, niter = 900. Finally, the performance of different recommendation algorithms is evaluated. It can be seen from experimental results that collaborative filtering travel route recommendation algorithm based KDE and classification is slightly higher than the route correlation rate of LDA travel route recommendation algorithm based KDE and classification by about 5%. In the actual application process, different recommendation algorithms can be selected according to users' actual demand. The final experimental results show that the optimization effect of proposed algorithm by using the classification method and KDE algorithm is obvious. The LDA and collaborative filtering algorithm optimized by classification method improves the route correlation rate and makes the route correlation rate indicator reach more than 90%. The KDE algorithm is used to optimize playing time, which makes playing time of cities more reasonable, which proves that the method of this paper has great reference value.

V. Conclusion

This paper proposed the travel route recommendation and generation algorithm based on LDA and collaborative filtering. The core of algorithm is LDA topic model and collaborative filtering. The LDA and collaborative filtering travel route recommendation algorithm based on KDE and classification are proposed in this paper. Although optimized algorithm designed has achieved good performance, but it still needs a lot of work to be done, including:

1) The recommendation algorithm based on LDA topic model has a certain degree of randomness in generating the recommended city list, there will be not related to the historical routes in resulting travel routes, but within the acceptable error rate. The output of recommendation algorithm based on collaborative filtering is relatively fixed and does not generate new feasible routes. And although the collaborative filtering algorithm does not have randomness problems, due to the irrelevance of user s' input city list, a certain error rate will also occur. Therefore, we can study a method that can combine the LDA topic model and collaborative filtering algorithm to make the performance of the recommendation algorithm better.

2) So far, the hyperparameters of LDA model, such as the number of topic k, α and β, are mainly adjusted manually by empirical rules, resulting in a huge amount of experimental work. Later, we can consider some methods of adding reinforcement learning and self-game, and propose a method that can learn the optimal parameters. This is also a research hotspot in the field of machine learning in recent years.

3) Further studying the evaluation method of travel route, because the evaluation of travel route has certain subjectivity, so this brings certain difficulties to actual assessment. At present, only quantifiable indicators can be extracted to evaluate part reasonability of travel route. So evaluation indicators may not be comprehensive. Later, we can study and propose a comprehensive and reasonable evaluation method of travel route.

Acknowledgements

This paper is partially supported by Qinghai University Student Science and Technology Innovation Fund Project (No. 2017-QX-12);

Reference

[1] Wohlin C, Runeson P, Höst M, et al. Experimentation in Software Engineering [J]. IEEE Transactions on Software Engineering, 2012, SE-12(7):733-743.

[2] Mabao Z, Research on Methods and Techniques of Tourism Travel Decision Support System [D]., Shandong University of Science and Technology. 54-58

[3] Baohui Jin. Travel route choice model based on regret theory. COMPUTER MODELLING & NEW TECHNOLOGIES 2014 18(4) 158-163

[4] Changing X, Kewen X, Yongwei Q, et al. Tourist route recommendation based on dynamic clustering [J]. Journal of Computer Applications, 2017, 37(8):2395-2400.

[5] Hui W, Changhua L, Yuling W ,et al. Application of ant colony optimization in Tourism Route Planning [J]. Software Guide, 2016, 15(4):144-146.

[6] Hou L, Yang H, Fan Y, et al. Research on Personalized Trip Itinerary Based on ILS-CS Optimization [J]. Journal of Frontiers of Computer Science & Technology, 2016, 10(1):142-150.

[7] Yajun Li ,Qing Lu, Changyong Liang .Review of Collaborative Filtering[J].Pattern Re-cognition and Artificial Intelligence,2014, 27(8):720-734.

[8] Qiang C, Dongmei H , Haisheng L ,et al . Personalized resource recommendations based on tags and collaborative filtering [J].Computer Science, 2014,41(1):69-71.

[9] Linden G, Smith B, York J. Amazon.com Recommendations: Item-to-Item Collaborative Filtering [J]. IEEE Internet Computing, 2003, 7(1):76-80.

[10] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2008, 3:993-1022.

[11] Krestel R, Fankhauser P, Nejdl W. Latent dirichlet allocation for tag recommendation[C].ACM Conference on Recommender Systems, Recsys 2009, New York, Ny, Usa, October. DBLP, 2009:61-68.

[12] Epanechnikov V A. Non-Parametric Estimation of a Multivariate Probability Density [J]. Theory of Probability & Its Applications, 2006, 14(1):153-158.

[13] Friedman N, Dan G, Goldszmidt M. Bayesian Network Classifiers [J]. Machine Learning, 1997, 29(2-3):131-163.

[14] Heinrich G. Parameter Estimation for Text Analysis[J]. Technical Report, 2008.

[15] Billsus D, Pazzani M J. Learning Collaborative Information Filters.[C]//Proceedings of the 15th National Conference on Artificial Intelligence (AAAI1998). San Francisco: AAAI Press, 1998: 46-54.

[16] BASU C, HIRSH H, and COHEN W. Recommendation as classification: using social and content-based information in recommendation[C]//Proceedings of the 15th National Conference on Artificial Intelligence. San Francisco: AAAI Press, 1998: 714-720.

[17] Yang S H, Long B, Smola A J, et al. Collaborative competitive filtering:learning recommender using context of user choice[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2011:295-304.

[18] George G, Haas M R, Pentland A. Big Data and Management: From the Editors [J]. Academy of Management Journal, 2014, 57(2):321-326.