Trip Recommendation Algorithm Based on Attraction Tags

Dupeng Wang\(^a\), Biyang Ma\(^b\), and Langcai Cao\(^c\)*

Dept. of Automation, Xiamen University, Xiamen 361102, P.R.China
\(^a\)1291721876@qq.com, \(^b\)biyangma@stu.xmu.edu.cn, \(^c\)langcai@xmu.edu.cn

**Keywords:** personalized recommendation; user interest; k-means algorithm

**Abstract:** Many route recommendation algorithms have been presented recently, it’s necessary to consider user’s preference in the recommendation. A travel recommendation algorithm is proposed based on visitor preferences. It analyzes the user's preference for different types of attractions and forms a user-preference matrix. Then it calculates an initial clustering center based on the interest distributed by K-means algorithm, and establishes a neighboring set for the target user to score the historical users’ route value for target user by target user’s reference on scenic spot type distribution. The method finds the historical user route with the largest value for the target user, thereby generating the trip recommendation. The experimental results show that the algorithm can quickly calculate the smaller neighboring users and obtain the recommended results. It not only has faster recommendation efficiency, but also has better recommendation accuracy. It provides a good service on personalized route recommendation.

1. Introduction

The booming tourism industry and the enthusiasm of mass tourism have made the massive tourism information provided by the current mainstream tourism information service platform overload, and an important method to solve the overload of tourism information is personalized recommendation technology [1]. There are several problems in the passing route recommendation: first, find some attractions that meet the interests of the target tourists and recommend them to tourists from a large number of scenic spots in a certain place, through some technical means (such as collaborative filtering). The mode ignores the distance of the attraction in the space and the user may not have enough time to visit all the attractions: Secondly, when judging whether the target user is interested in an attraction, we often rely on other tourists to evaluate the attraction, but people with different interests and hobbies often have different evaluations of the same attraction. In today's valuable information data, it is very difficult to obtain similar hobbies and historical user information. Third, constructing an optimal solution requires all permutations of the scenic spots to be computed, which is an NP-hard problem. The computational cost will be prohibitive even for a small number of scenic spots.

There are many research results on personalized travel recommendation at home and abroad. The existing algorithms can be divided into two categories: one is to obtain users’ requirements through question and answer interaction, and to generate recommendations for attractions. The other is to...
get user interest preferences by mining users’ history logs. To generate recommendations for travel information, most of the current algorithms rely on the users’ rating information on attractions and travel routes to match "similar users" to achieve travel recommendations.

![Figure 1 Example of Similar Users’ Diagram.](image)

In social networks, people with similar hobbies are often the same type, and their evaluation of attractions is more informative in this small social circle: So visitors often select tourists with similar interests when referring to the historical evaluation of scenic spots. This paper builds a user model based on the theoretical basis that the route choice is influenced by user’s preference for scenic spots. The user optimizes the initial clustering center of k-means according to the user's preference for multiple scenic spot types, and introduces the calculation method of route evaluation score. The route that satisfies the constraint is recommended to the user.

2. Traditional Route Recommendation Algorithm

The core idea of the route recommendation algorithm is to obtain the closest user set of the target user through evaluation and analysis of the user’s interest, and then predict the target user's interest in the project according to the evaluation of the project by the nearest neighbor user set, so we generate a recommendation [3]. The collaborative filtering algorithm is widely used in personalized recommendation technology, and the technology is relatively mature. According to the statement, in the collaborative filtering algorithm, the input initial data can usually be expressed as an M×N user-score matrix. In this matrix, M is the number of users, N is the number of attractions, and the value in the matrix can be considered to be the rating of user M for attraction N. The proximity matrix is established, and after acquiring the user-attribute rating matrix data, a similar user having the same interest as the target user is searched. However, the traditional collaborative filtering algorithm requires too much data information, and cannot produce satisfactory results in the case of insufficient sample information.

The traditional recommendation algorithm is often very simple in the analysis of the scores of the scenic spots. It only considers the score of the scenic spots, and does not consider the scores of the scenic spots on the type attributes. For example, an attraction includes history and religion. Considering the difference in scores of different people in the same attraction, a tourist who does not like historical attractions and a tourist who likes historical attractions have a large difference in scores for a historical attraction. For those who like historical attractions, they also like historical visitors are more informative about the attraction, and visitors who do not like historical attractions can ignore it. The focus of this paper is to recommend a route that matches the target visitors in the path of historical users with similar interests, and this route is the most popular.
3. Trip Recommendation Algorithm Based on K-means

3.1. Clustering Algorithm

Cluster analysis [5] divides data into meaningful or useful groups, which could capture the natural structure of the data. By clustering grouping, objects with higher similarity are grouped into the same group, and the direct similarity of different groups is different [6]. The higher the degree of similarity within a group, the greater the difference between groups and the better the clustering effect. From the perspective of practical application, clustering analysis is one of the main tasks of data mining. Moreover, clustering can be used as an independent tool to obtain the distribution of data, observe the characteristics of each cluster of data, and focus on the further analysis of a particular cluster of clusters. Clustering analysis can also be used as a preprocessing step for other algorithms, such as classification and qualitative induction algorithms.

Algorithm 1: K-MEANS

1. Select \( K \) points as the initial centroid.
2. Repeat:
3. Assign each point to the nearest centroid, forming \( K \) clusters
4. Recalculate the centroid of each cluster
5. Until the centroid no longer changes

3.2. Algorithm Flow

Through research and analysis of existing tourist user data, it is concluded that there are five main types of scenic spots: natural landscape, historical and cultural, folk customs, religious holy land and entertainment shopping. The algorithm firstly obtains the preference of the tourist users for these types of attractions through data mining technology and interest distribution calculation. Then we establish the user-attraction distribution matrix. We use a variety of scenic spot types as the initial clustering of K-means, and the tourist users with higher similarity of attraction types in the clustering route form a set of nearest neighboring tourist. The average attribute scores of the different attributes of the users of each cluster are calculated, and the attribute scores of each attraction are obtained. Finally the route evaluation of the target tourists is used. The value is calculated and the one with the highest score is recommended to the visitor. The algorithm flow is shown in Figure 2.

![Figure 2 Algorithm Flow](image-url)
4. Algorithm Framework

4.1. Modeling of Travel User Destination Preferences

The user model is the basis for the realization of personalized travel itinerary. This paper uses the space vector model based on the type of tourist attractions to represent the user's preference for the type of attraction, that is, the model is represented as an $N \times 5$ User–tour destination type preference matrix $SNM$, where $N$ represents the number of users, $M$ represents the category 5 tourist attractions, and the $SNM$ value represents the number of the $M$ attraction types in the $N$th user history route. This matrix can be explained in Table 1.

<table>
<thead>
<tr>
<th>User</th>
<th>Nature</th>
<th>Shopping</th>
<th>Historical</th>
<th>Religion</th>
<th>Customs</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>U2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>U3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>U4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>Un</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

The preference of a tourist destination is defined as: counting the scenic spots of the travel selected by the travel user in a certain period of time, and the preference of the scenic spot type is the ratio of the number of scenic spots included in the traveled route selected by the user to the total number of scenic spots in the route [4]:

$$C_i = \frac{m_i}{\sum_{n=1}^{n} m}$$ (1)

In Equation (1), $n$ is the total number of attractions in a historical user route; $m$ is the number of attractions of a certain type in the route. If the historical user route does not contain a certain attraction type, the user's preference for the type of attraction is 0.

4.2. K-means Algorithm Based on Attraction Type Preference

The core of the algorithm is that it can randomly generate cluster centers and unsupervised classification of target groups. However, the randomly selected cluster centers cannot reflect the preference differences of various types of attractions. Therefore, we set five initial cluster centers through the distribution of five types of scenic spots obtained from the previous analysis, and cluster all tourist users into five nearest neighbors. The distribution of attractions in the user history route of each cluster class is generally similar, and the user interests of each cluster class can be considered similar. The target user is found to belong to the cluster class by calculating the difference between the target user and each cluster class center. We use the cosine distance metric [7] to represent the similarity between texts, which defines the similarity of two routes through Equation (2).

$$\text{sim}(r_i, r_j) = \frac{\sum_{k=1}^{n} w_k(r_i) \cdot w_k(r_j)}{\sqrt{\sum_{k=1}^{n} w_k(r_i)^2} \cdot \sqrt{\sum_{k=1}^{n} w_k(r_j)^2}}$$ (2)

$w_k(r_i)$ is the proportion of the route of the user $i$ in the $k$ type of attractions.
4.3. Route scoring Algorithm

After the user obtains the nearest neighbor set, the best route calculation method becomes an important means for the target user to find the best route, and is also a key step of the user route [8,9,10]. This paper analyzes the route scoring algorithm and calculates the user’s score on a historical user route.

Tourist routes \( R =< v_1, v_2, v_3, ..., v_n > \), \( v_n \) represents the scenic spots passing through the route; each attraction \( V \) can contain multiple types of attributes, such as an attraction that can contain two types of attributes: historical culture and religious holy land, namely: \( K = \{ k_1, k_2 \} \). \( K = \{ k_1, k_2, ..., k_q \} \) as the type of attraction, \( \lambda_{kq} (> 0) \) is the size of the tourist’s interest in the type of attraction. We get the target visitors’ total score for a route as Equation (3).

\[
KC(R) = \sum_{kq \in K} \lambda_{kq} cov_{kq}(R)
\]  

(3)

\( cov_{kq}(R) \) is the value of the representative \( K_q \) type of attraction in the historical tourist route as Equation (4).

\[
cov_{kq}(R) = 1 - \prod_{v \in R} [1 - \text{cov}_{kq}(v)]
\]

(4)

For example, a route \( R_1 =< v_1, v_2, v_3, v_4, v_5 > \), \( v_1 \) (Historical: 0.7; Shopping: 0.1), \( v_2 \) (Nature: 0.5), \( v_3 \) (Nature: 0.6), \( v_4 \) (Historical: 0.5, Religion: 0.8), \( v_5 \) (Shopping: 0.3, Customs: 0.5). If the target tourist’s interest rate for these five types (Nature: Shopping: Historical: Religion: Customs) of attractions is 5:4:3:1:0; then the total score obtained by this route can be calculated by:

\[
KC(R_1) = 5 \times [1 - (1 - 0.5)(1 - 0.6)] + 4 \times [1 - (1 - 0.1)(1 - 0.3)] + 3 \times [1 - 0.7 \times 0.5] + 1 \times [1 - (1 - 0.8)] = 8.83.
\]

In this paper, \( cov_{kq}(R) \) is the value of the representative \( K_q \) type of attraction in the historical tourist route, through Equation (5).

\[
cov_{kq}(v_i) = \frac{\sum_{j=1}^{n} k_{qj}}{n}
\]

(5)

4.4. Producing Recommendations

In order not to make the recommendation result too complicated, we try to recommend the historical route of similar attraction type distribution to the user according to the target user's choice of the type of attraction type, that is, obtain the one with the largest score by formula 6, and formula 7 indicates that the route meets the tourists. The time constraints should not exceed the travel time specified by the visitor’s plan.

\[
R = \arg \max_R KC(R)
\]

(6)

\[
S. t. BS (R) \leq \triangle
\]

(7)

5. Experimental Results and Analysis

5.1. Experimental Data

We select travel users who have more than 70 travel notes or more than 10 questions from the Raiders community of Apache and trip. Through text mining and user analysis, we randomly extract 10 travel itineraries for each user and determine the types of travel destinations that users have chosen (Table 2) The experiment uses java language, based on eclipse development platform, and
uses SQL database technology to build users and data storage.
The travel route information of a user is: Attractions #Type# Rating

5.2. Experimental Methods

Firstly, according to the algorithm flow and experimental data, the preference of each user for the type of attraction is calculated, and the sum of the preferences is 1. Then the user clustering is performed to obtain the 5 types of nearest neighbors with the tourist attraction type preference as the cluster center (Table 2). Finally, we obtain the target user's requirements for the distribution of recommended route attractions, and then calculate the difference between the cluster center and the target user of each type of nearest user set to find the user set to which the target user belongs, and ensure the target user and other users in the user set. The types of attractions are similarly distributed.

Table 2 User - Attraction Type Preference Matrix.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Nature</th>
<th>Shopping</th>
<th>Historical</th>
<th>Religion</th>
<th>Customs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1071</td>
<td>0.1786</td>
<td>0.1071</td>
<td>0.2857</td>
<td>0.3214</td>
</tr>
<tr>
<td>2</td>
<td>0.3684</td>
<td>0.1579</td>
<td>0.3158</td>
<td>0.0526</td>
<td>0.1053</td>
</tr>
<tr>
<td>3</td>
<td>0.0833</td>
<td>0.2083</td>
<td>0.2917</td>
<td>0.3750</td>
<td>0.0417</td>
</tr>
<tr>
<td>4</td>
<td>0.3684</td>
<td>0.1579</td>
<td>0.1053</td>
<td>0.2105</td>
<td>0.1579</td>
</tr>
<tr>
<td>5</td>
<td>0.2500</td>
<td>0.3750</td>
<td>0.1875</td>
<td>0.0625</td>
<td>0.1250</td>
</tr>
<tr>
<td>6</td>
<td>0.1429</td>
<td>0.2143</td>
<td>0.0714</td>
<td>0.1429</td>
<td>0.4286</td>
</tr>
</tbody>
</table>

5.3. Experimental Results and Analysis

When recommending a route for a visitor, first obtain a distribution of tourist attraction type hobbies, which can be obtained by scoring various types. Such as nature class 4 Shopping class 2 Historical class 1 Religion class 3 Customs class 5. We can get recommended routes (Figure 3) under different time constraints.

Table 3 User - Attraction Type Preference Matrix.

<table>
<thead>
<tr>
<th>Time</th>
<th>Score</th>
<th>Sum(Scenic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T&lt;4h</td>
<td>3.42</td>
<td>4</td>
</tr>
<tr>
<td>T&lt;6h</td>
<td>5.76</td>
<td>6</td>
</tr>
<tr>
<td>T&lt;8h</td>
<td>6.96</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 3 A case study showing the trips planned using different time.
We used the Internet to send 140 travel recommendation lists to the original experiment users and returned 116 results. Then we randomly selected 100 results, and used the accuracy evaluation algorithm and the satisfaction survey. (If the first recommendation is adopted, the satisfaction is very good; if the second recommendation is adopted, it is recorded as good; if the third recommendation is adopted, it is recorded as general; if the recommendation is not taken, the satisfaction is recorded as poor) to test the recommendation result.

\[
\text{Precision} = \frac{RC}{RC+RU}
\]  

Among Equation (8), RC is the recommended number of trips adopted by the user; RU is the recommended number of trips not adopted; RC+RU is 100. The experimental results show that the recommended accuracy rate (user adoption rate) is 87.4%. With a good score of 10 points, a good score of 8 points, a general score of 6 points, and a score of 0 points, the cumulative user satisfaction value is 806 (total score 1 000), that is the satisfaction rate is 80.6%.

6. Conclusion

This paper proposes a travel itinerary recommendation algorithm based on K-means clustering by studying the personalized recommendation problem of travel itinerary. The algorithm calculates the user's preference for the type of travel destination and uses it instead of the project score to calculate the similarity between users, making it suitable for personalized recommendations under the user's multi-interest. Experiments show that the algorithm has higher accuracy and recommendation efficiency, and it can improve the satisfaction of personalized travel itinerary recommendation. However, the current tourism preference model is still not perfect, and the interest modeling of tourism users’ needs further research. In the next step, we can start with user travel motivation and get more user modeling methods to seek better recommendations.

References