Intelligent Tourism Recommendation System

z5457836 Zhenxuan Fang

1.Scope

Domain: Intelligent recommendation of tourist destinations and personalized itinerary planning.

In this project, I will conduct a comprehensive analysis of user reviews from TripAdvisor, check-in data from Foursquare, and hotel reservation records. By integrating users' interests, past behaviors, and geographical locations, I will design an intelligent recommendation system to provide personalized tourism attractions and itinerary planning. The recommended content of this system not only includes the best tourist destinations and popular attractions but also recommends suitable hotel accommodation options based on users' needs. At the same time, the system customizes complete travel routes according to users' interests and specific requirements to help users plan their trips more efficiently and conveniently. Through this system, I can enhance users' travel experience and perfectly meet their needs.

Intended users:

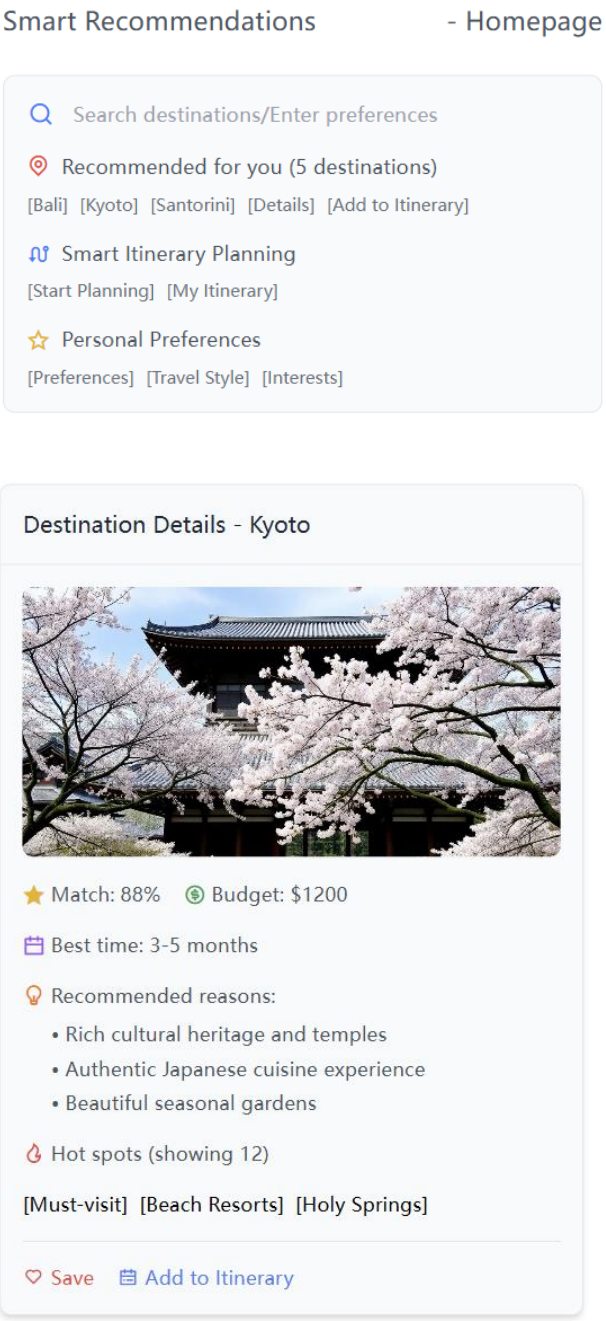
Individual travelers: People aged 20 to 50, passionate tourists with purchasing power

Family members: Planning family trips considering each person's preferences

Business travelers: Considering the need for efficient itinerary arrangement during the travel period

The characteristics of the users are that they travel more than twice a year on average, and they highly value personalized experiences as well as the arrangement of time. At the same time, they are willing to spend a certain amount of money for this service.

The design of the user interface: It is achieved through mobile. Below is an example of an interface that I made. Each time, five major destinations are presented. For the recommendations of attractions, it would be 8 to 10. Regarding hotels, three options of different price ranges are provided. Finally, the itinerary planning involves generating a detailed 3 to 5 day schedule.



User interaction: This system will collect the information demonstrated by users through explicit feedback and implicit feedback. In the feedback display, users can rate the experience of the attractions and itineraries, and provide brief comments to help the system understand their preferences. At the same time, users can also set the tags they are interested in to enable the system to push more precisely the content that matches their interests. In implicit feedback, we can understand what the user is usually interested in through their search keywords, filtering conditions and the content in their favorites. We can also make this judgment based on the user's duration of stay and whether they click to read more details. After collecting all these data, the system will conduct real-time learning. It will update the weights of users' personal preferences through daily

updates based on each user's interaction, in order to ensure that it can recommend content that users are interested in and maintain timeliness.

Solution to the cold start problem: In the aspect of new user cold start, information such as the user's age, location, and occupation will be used. At the same time, upon entering the system, there will be a preference questionnaire on the interface. Through the classification of several major tabs, the profile of new users can be quickly obtained. Users can also choose their desired destinations by referring to the popular travel lists. In terms of the cold start process for new destinations, a label will be created based on the geographical location, climate and culture of the destination, as well as the unique features of each place. Then, recommendations will be made based on this label.

Business model design: The revenue source is to collect a 5% commission from the reservations of cooperative hotels, attraction tickets, or similar transportation services such as car rentals. Additionally, through software subscription, a fixed monthly fee can be charged to enjoy personalized travel recommendations. At the same time, advertising fees from attractions and other organizations can also be collected.

2. Datasets

1. Selection of the main dataset:

1）TripAdvisor Hotel Review Dataset
https://www.kaggle.com/datasets/joebeachcapital/hotel-reviews

Key fields and application of recommendation systems:

reviewer_id-- User unique identifier（Constructing the User-Item Interaction Matrix, the Foundation of Collaborative Filtering）

hotel_name-- Hotel Name（Item identification, used for constructing the recommendation target set）

reviewer_score-- User rating（Explicit feedback signal, training score prediction model）

review_text-- Commentary text content（Emotional analysis and text mining, extracting users' fine-grained preferences）

review_date-- Commenting time（Temporal modeling, capturing changes in user preferences and seasonal patterns）

hotel_address-- Hotel Address（Geographical location characteristics, supporting

location-based recommendations）

additional_number_of_scoring-- Total number of comments（Item popularity characteristics, handling popularity deviations in recommendations）

2）Foursquare's global check-in data set

https://sites.google.com/site/yangdingqi/home/foursquare-dataset

Key fields and application of recommendation systems:

user_id-- User identification（User behavior sequence modeling, learning user movement patterns）

venue_id-- Location Identifier（POI recommended target items）

latitude, longitude-- GPS coordinates（Geographical location clustering, spatial similarity calculation）

utc_time-- Check-in time（Temporal recommendation, learning users' activity time preferences）

venue_category-- Place category（An important feature of content filtering, user interest classification）

country, city-- Geographical location（Hierarchical geographic recommendation）

3）Expedia Hotel Recommendation Competition Dataset

https://www.kaggle.com/c/expedia-hotel-recommendations

Key fields and application of recommendation systems:

user_id-- User identification（User profile construction）

srch_destination_id-- Search destination ID（Destination recommendation）

hotel_cluster (0-99)-- Hotel clustering labels（Recommendation target）

is_booking-- Is it an actual reservation (0/1)? （Core tags for conversion rate prediction）

hotel_country, hotel_market-- Hotel geographic information（Geographic preference modeling）

srch_adults_cnt, srch_children_cnt-- Number of searchers（Preference distinction between family travel and individual travel）

srch_ci, srch_co-- Check-in and check-out dates（Preference for travel duration,

seasonal analysis)

2. Data set application strategy

In this project, a hierarchical data utilization strategy can be adopted. In simple terms, it involves integrating multiple data sets to optimize the performance of the tourism recommendation system. It is divided into three layers. The first layer is to identify the users' actual search and booking behaviors through the Expedia data set to provide implicit feedback. This is applied to construct the user-item basic interaction matrix and provide training data for the collaborative filtering algorithm. The second layer is to provide sufficient text and ratings through the TripAdvisor data set to depict the user's profile, and then add an emotional dimension to the recommendation model through the method of sentiment analysis. The third layer is to provide the users' geographical location and mobility preferences through the Foursquare data set to support the spatial recommendation algorithm and POI (Point of Interest) discovery, helping to improve and develop location-based recommendation and path planning algorithms. For the specific recommendation tasks, they can be divided into four tasks: The first one is to conduct destination recommendation by training a collaborative filtering and geographic clustering model using the destination search data from Expedia and the city check-in data from Foursquare, while also considering the user's historical preferences and geographical similarity. The second one is to recommend hotels by using a sentiment-enhanced hybrid recommendation algorithm that considers the user's price sensitivity, service quality, and geographical convenience based on TripAdvisor reviews and Expedia booking behaviors. The third one is to recommend POIs (Points of Interest) by using content filtering and geographic perception recommendation with the check-in data from Foursquare and the attraction comments from TripAdvisor, combined with the user's interests and geographical distance. The fourth one is to solve the personalized itinerary planning problem under time and geographical constraints by using reinforcement learning and genetic algorithm optimization methods, in order to address the user's personalized itinerary.

3. Analysis of Dataset Limitations and Solutions

There is a bias in the TripAdvisor dataset regarding the user group. Most of the reviews come from users with extreme experiences, while there are insufficient reviews from users with moderate experiences. By using social media or online data for statistics, a broader range of user preferences can be obtained. The data in the Foursquare dataset was collected from 2012 to 2013 and lacks timeliness, thus failing to reflect current user behaviors. However, most of the attractions are well-known ones and do not change much. But it should be updated based on the current real-time situation. At the same

time, more new datasets can be sought to improve this part. In the Expedia dataset, most of the users are of middle to high-end consumption level, and there is a relatively small amount of data for budget-friendly travel users. Data can be obtained by using an initial online questionnaire and combining certain incentive mechanisms to encourage users to participate in the survey. All of the above datasets have issues of cleanliness, such as fake negative reviews, fake positive reviews, or data missing. These problems can be solved by using Bayesian methods to handle data uncertainty and reduce the influence of false data, and by inferring the missing profile information based on user behavior patterns.

Specifically, I will use the surprise library to address the issue of false data, employing various recommendation algorithms such as SVD and KNN. Additionally, cross-validation can be utilized to mitigate data noise and enhance the stability of the model. Furthermore, by using the weighted scoring method, we can reduce the influence of extreme users on the model to solve the problem of extreme users. Then, the scikit-learn library is used to address the issues of data missing values and outlier detection, as scikit-learn offers a variety of preprocessing tools, such as data filling, standardization, and normalization. And the algorithms used for outlier detection, such as Isolation Forest and LOF. Then, the torchrec library is used to establish nonlinear relationships through deep learning models, which can easily respond to the noise in the data and solve the problem of data cleanliness.

3. Methods

In this system, a multi-level hybrid recommendation system is designed, integrating three methods: emotion-enhanced collaborative filtering, context-aware geographic recommendation, and sequential itinerary planning recommendation.

The basic approach is the emotion-enhanced collaborative filtering recommendation method. This method was chosen because the decision-making of travel plans is entirely dependent on the user's emotional experience and the sharing of experiences by users similar to them. And collaborative filtering is precisely capable of efficiently identifying the preference patterns of user groups. At the same time, through sentiment analysis, more comprehensive preference information can be extracted from the text of user comments. Compared with the traditional rating system, it can better capture users' preferences. It effectively solves the problem of sparsity in the processing of tourism data and can also capture users' potential preferences. The core of the technology is to use the BERT model to extract the emotional features in the comments, and then integrate this as additional information into the SVD++ algorithm to enhance the effect of collaborative filtering. Then, the similarity of ratings and emotions is

combined to calculate the similarity of users.

The second method is context-aware geographic recommendation. This method is chosen because tourism recommendations have very crucial geographical attributes and seasonality. Traditional collaborative filtering methods do not take these factors into account. This method can perform geographical clustering and time series modeling, thereby improving the accuracy and satisfaction of recommendations. The core of implementation is to use the HDBSCAN algorithm to cluster user check-in data geographically, and use geographical proximity as the weight to adjust the similarity calculation. Then, the user's comments and check-in time are used to determine the frequency of the user's travel.

The third method is serialized itinerary planning recommendation. As is known to all, tourism is a series of connected steps. The itinerary planning of tourism has a very strong logical dependence. Every user's choice will affect the subsequent itinerary. The user's destination, hotel, and visited attractions are all interrelated. Serialization can precisely help optimize the overall practicality of the itinerary. The core is to use RNN or Transformer to build a model of the user's historical travel sequence, learning the changes in the user's historical decisions and preferences. Then, the genetic algorithm is used to solve the constrained itinerary planning problem, and finally, the itinerary planning is modeled as a Markov decision process for optimization.

Considering the different limitations of the three methods, a hybrid recommendation system integration strategy is adopted. The results of emotional-enhanced collaborative filtering geographic recommendation and serialized itinerary planning recommendation are fused through a linear weighted fusion method to improve the accuracy of recommendations.
Finally, the effectiveness of each method is evaluated by comparing the results after incorporating these methods with the base method, using metrics such as hit rate and recall@k.

4. Evaluation

1) Recommendation Model Assessment: Based on historical data, the following indicators are used to evaluate the performance of the model. Among them, the Top-N recommendation metric is the core dimension for assessment. Precision@5 measures the proportion of relevant items included in the recommended list, which directly reflects the accuracy of the recommendation. Recall@5 can evaluate the extent to which the recommended items cover the users' true interests. NDCG@5 considers the weights of the recommended positions to assess the quality of the ranking. Additionally, there is

the Coverage metric, which is used to measure the coverage ability of the recommendation system for less popular destinations, which is crucial for the diversity of tourism recommendations.

2) Recommender System Assessment: The click-through rate of any application is the most direct user feedback, which can be used to determine what content users are satisfied with. Also, the duration of users' stay on the page, the rate of their favorites, and the sharing rate all indicate the degree of users' interest and approval for the recommended content. There is also the conversion rate indicator, which tracks the conversion rate of users from the recommendation stage to actual booking. This directly reflects the commercial value of the tourism recommendation system. In addition, there are questionnaires for user satisfaction to determine whether users feel that the recommended content meets their expectations, whether a diverse range of options is provided, and their level of satisfaction with the service.

3) Multi-index trade-off and model selection: Among these indicators, the most important ones are the user conversion rate and satisfaction level. These two indicators can directly reflect the actual value brought by the recommendation system. The second important metric is Precision@K and NDCG@5, which are indicators reflecting the accuracy of the model. Regarding how to select the best model based on indicators, I believe the main issue is to address the trade-off between Precision and Coverage. This can be achieved through dynamic adjustment, where the weights can be adjusted based on whether the users are new or old. Such continuous dynamic optimization can lead to the best model.

4) Computing requirements and considerations for real-time performance: The computing requirements of the system can be divided into two parts: offline training and online recommendation. During the offline stage, by conducting real-time updates to the model, since user preferences may change in different situations, an incremental learning strategy can be adopted to update user behavior characteristics and other related items on a daily or weekly basis. Then, during the online recommendation stage, the results are recommended to the users through the neighbor recall mechanism. At the same time, a standard can be set that the training time of the model offline should not exceed 1 hour, and the response time for online recommendations should be within 300ms.

5) User research design:

The content of the questionnaire is as follows (with a scoring system of 1-5, where 5 represents the highest score):

How well does the recommended content match your travel preferences? (1-5)

Did the recommendation results offer diversity and personalization? (1-5)

Would you choose the products related to your reservation by using this recommendation system? (1-5)

What are the shortcomings of this system?

These questionnaires will serve as a reference for the subsequent improvement of the recommendation system.