Tourism recommendation system based on data mining

Zehao Wang¹, Bing Liu² and Bing Liu¹*

¹Geomatics, Shandong University of Science and Technology, Qingdao, Shandong, 266590, China
²Geomatics, Shandong University of Science and Technology, Qingdao, Shandong, 266590, China
*Corresponding author’s e-mail: lb@sdust.edu.cn

Abstract. With the ever-increasing living standards of the people in recent years, more and more people have entered the army of tourism, and our country’s tourism industry has achieved unprecedented development. A series of travel portal websites such as Ctrip.com, Qunar, and Malacca have emerged. A large amount of tourist information is presented to users, but at the same time it has led to the blind choice of users. Massive data have overwhelmed the information that users are really interested in. The emergence of tourism recommendation systems helps users solve this problem. Against the above background, this paper designs and implements a tourism recommendation system based on data mining. From the perspective of mining the similarity between users, the similarity between users is calculated through the collaborative filtering algorithm, and then the attractions visited by users with higher similarity are recommended.

1. Introduction

With the further deepening of China's comprehensive construction of a well-off society, China's tourism industry will usher in a positive opportunity for development: the domestic economy will increase steadily, which will generate strong tourism demand; the disposable income of residents will increase steadily, and by 2020, per capita GDP may reach around 3,500 US dollars or more, and it will usher in the explosive growth phase of tourism; China's social stability is the safest tourist destination in the world; and the improvement of the vacation system will further promote the development of the tourism market [1].

On the other hand, due to the rapid development of the Internet and the explosive growth of network information, a large number of tourist user websites have emerged, which has won unanimous praise from users. Not only does it collect and integrate a large amount of tourist information, but also obtains a large number of data from travel website users. From the perspective of the user, the user cannot quickly and accurately find the information he needs from the massive data, which creates an "information overload" problem. If you want to solve the information overload problem, you may need the popular personalized recommendation technology. The personalized recommendation system is developed from e-commerce. Based on massive data mining, it provides users with personalized product recommendation and decision support. However, the application of such personalized recommendations on domestic travel portals is not very extensive.

Based on the above background, this paper designs and implements a travel recommendation APP based on data mining, including popular attraction ranking based on Baidu search list, attraction
recommendation based on collaborative filtering algorithm, and route planning and navigation functions.

2. Key technologies for system implementation

2.1. Web Crawler
Web crawlers (also known as web spiders, web robots) are programs or scripts that automatically crawl web information according to certain rules [2].

Web crawler is a program that automatically extracts web pages. It downloads web pages from the World Wide Web for search engines and is an important component of search engines. It can be divided into two types: general reptiles and focused reptiles. This article mainly uses focused reptiles. The workflow of focusing on crawlers is more complicated. It is necessary to filter the links irrelevant to the topic according to certain webpage analysis algorithms, retain useful links and put them into the URL queue waiting for crawling [3]. Then, it will select the URL of the web page to be crawled from the queue according to a certain search strategy, and repeat the above process until it reaches a certain condition of the system. In addition, all web pages crawled by the crawler will be stored by the system, analyzed, filtered, and indexed for later query and retrieval. For focused crawlers, the analysis results obtained by this process may also be Give feedback and guidance on the subsequent crawling process.

Focus reptile is a program that automatically downloads web pages. It selectively accesses web pages on the World Wide Web and related links to obtain the required information according to the established crawl target. Unlike the general purpose web crawler, the focus crawler does not pursue large coverage, but targets the crawling of web pages related to a particular topic content to prepare data resources for topic-oriented user queries [4].

2.2. Collaborative filtering algorithm
The collaborative filtering recommendation algorithm is the earliest and more famous recommendation algorithm. The main function is prediction and recommendation. The algorithm discovers the user's preferences by mining the user's historical behavior data, and classifies the users based on different preferences and recommends similar products with similar tastes [5]. The collaborative filtering recommendation algorithm is divided into two categories, namely, a user-based collaborative filtering algorithm and an item-based collaborative filtering algorithm. To put it simply, people gather together and things are divided into groups. We use an item-based collaborative filtering algorithm.

The item-based collaborative filtering algorithm is very similar to the user-based collaborative filtering algorithm, which interchanges goods and users. The relationship between items is obtained by calculating the scores of different users for different items. Recommending similar items to users based on the relationship between items. The ratings here represent the user's attitudes and preferences for the product. To put it simply, if user A purchases both item 1 and item 2, then the correlation between item 1 and item 2 is high. When User B also purchased the item 1, it can be inferred that he also has the demand to purchase the item 2 [7].

Regarding the calculation of similarity, several existing basic methods are based on vector (Vector), which is to calculate the distance between two vectors. The closer the distance is, the greater the similarity [8]. In the recommended scenario, in the two-dimensional matrix of user-item preferences, we can use a user's preference for all items as a vector to calculate the similarity between users, or to treat all users' preferences for an item. A vector to calculate the similarity between items. When calculating the similarity in this paper, the cosine similarity is used. The calculation method of cosine similarity is introduced below.

The cosine similarity uses the cosine of the angles of the two vectors in the vector space as the measure of the difference between the two individuals. The closer the cosine value is to 1, the closer the angle is to 0 degrees, that is, the more similar the two vectors are, this is called "cosine similarity" [9].
\[
\cos \frac{a \times b}{\|a\| \|b\|}
\]
\[
\cos \frac{x_1x_2 + y_1y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}
\]

2.3. Cold start problem

Cold start in the recommendation system indicates that the system accumulates too little data and cannot give personalized recommendations to new users. This is a big problem in product recommendation [10]. Basically, the cold start problem can be divided into the following three categories:

User cold start: User cold start mainly solves the problem of how to make personalized recommendations for new users. When a new user arrives, we don't have his behavioral data, so we can't predict its interest based on his historical behavior, so I can't give him a personalized recommendation.

Item Cold Start: The item cold start primarily addresses the issue of how to recommend new items to users who might be interested in it.

System cold start: The system cold start mainly solves how to design a personalized recommendation system on a newly developed website (no user, no user behavior, only some items of information), so that the user can experience the website when it is just released. Personalized recommendation service [11].

It is recommended that the system cold start problem is mainly caused by the lack of scalability of the system itself. In many cases in the industry, it is recommended by manual labeling. After getting enough feedback, run the appropriate algorithm online for testing, and then run it online. In this process, the category and keyword between the articles can be fully utilized for similarity calculation, and then the user's registration information and preliminary access records are used to make appropriate recommendations, forming a good closed loop.

3. Overall system design

3.1. Database Design

The database of the system uses the SQLite database, mainly related to the user's personal information, travel information, attraction search ranking information, similarity ranking information, etc. The main table information is listed below.

| Field        | Type  | Description         |
|--------------|-------|---------------------|
| ID           | Int   | Number              |
| User Name    | String| username            |
| User Pwd     | String| user password       |
| User Destination | String | destination       |
| User Category | String | Type of travel     |
| User Days    | Int   | Travel days         |
| User Cost    | Int   | User's cost         |

| Field       | Type  | Description       |
|-------------|-------|-------------------|
| Destination | String| destination       |
| Category    | String| Type of travel    |
| Days        | Int   | Travel days       |
| Costs       | Int   | Per capita spending|
| Scenery     | String| Tourist attraction |
3.2. Functional structure design

System function structure design refers to the design of the relationship between various functional modules of the system. Figure 1 lists the functional structure of the whole system:

![System function structure](image)

The function modules are described as follows:

1. User registration. Provide user registration and login function, support user password modification and remember password function, convenient for user input, and password reset function to prevent users from forgetting password, in order to reset password.

2. Information entry. Including user personal information entry and travel travel information entry, the former contains the destination the user wants to travel, the type of travel, the number of travel days and the per capita cost, the latter contains the city the user has traveled, the number of days traveled, the type, the cost and the tour. Attractions. This information is used when performing user similarity calculations.

3. Popular attractions. The Baidu Search Billboard Domestic Attractions module provides real-time search popularity rankings for domestic attractions. You can host the crawler program on the cloud platform and then call it in the Android background to get the returned json data, and then parse the obtained json data. To the user.

4. Recommended attractions. Through the user-based collaborative filtering algorithm, the similarity between the current user and all users in the database is calculated, and the scenic spots visited by the users with higher similarity are recommended to the current user, thereby saving the time for the user to find the interesting attractions.

5. Route planning. After obtaining the recommended tourist attractions, you can use the POI search function to obtain the geographical coordinates of the attractions, and then use the high-
definition map SDK for Android to make route planning based on these coordinate data, and call the navigation function [12].

4. Tourism recommendation system implementation

4.1. System main interface
The system interface design follows the dual principles of practicality and operability, aiming at creating a friendly and popular user interface. We analyze the interface of various software and combine the actual situation to develop the interface.

The main interface design is implemented by means of Activity Managed Fragment, similar to the Tab interface layout, similar to the common WeChat and QQ main interfaces.

Use Fragment to manage the layout and function implementation of the page corresponding to each tab. Then associate the Fragment with Android, so that Android only needs to manage the Fragment, it plays the role of the scheduler, no longer cares about the content and function implementation of each Fragment. This greatly liberates the Activity, making the code simple and easy to read. It can be seen that the Fragment is used to realize the page separation corresponding to the Activity and the Tab, especially when the layout and logic of the Tab corresponding page are more complicated, the benefits of using Fragment can be better understood [13].

4.2. Implementation of the main functions of the system

4.2.1. Popular attractions ranking
The module is mainly to show the current ranking of popular attractions, the data comes from the Baidu search list of domestic attractions search rankings. First, use the Python language to write a simplified version of the crawler. Compared to the full version of the crawler, the link discovery module is missing, because the data is all from a single web page; then the debugged program is hosted on the Marksman cloud platform, designed as a API application for real-time calls on the mobile side. The result returned is a json data set, which can be parsed after parsing the json data [14].

Because of the background data download process, the relationship between the thread and the main thread is involved. A thread is a single execution sequence. The code in a single thread is executed step by step. All Android apps run from the main thread. However, the main thread is not a predetermined execution sequence like a thread. Instead, it is in an infinite loop running state, waiting for a user or system to trigger an event. After the event is triggered, the main thread is responsible for executing the code in response to these events.

Network connection takes time. The web server may take 1-2 seconds to respond to the access request, and the file download takes longer. With this in mind, Android prohibits any main thread network connection behavior. Even if it is strong, Android will throw a Network On Main Thread Exception. Therefore, data downloads are often placed in a background thread [15].

At the same time, the module will display the distribution of hot spots in China through the form of heat map, which can help users to more intuitively feel the distribution of popular attractions in the country.

4.2.2. Recommended tourist attractions
The module uses a user-based collaborative filtering algorithm to calculate the similarity between the current logged-in user and all users in the database. The similarity is calculated by the cosine similarity method. When the similarity between all users is calculated, the re-sorting is performed, and then the attraction recommendation is given in order of similarity.

4.2.3. Route planning and navigation
When the user gets the recommended attraction, the next step is to plan the route. The route planning function is developed through the high-tech map SDK. Before development, you need to apply for the
key first, then create a new application in the console, bind the service, and also need the security code SHA1 and the corresponding Package Name.

After the route planning is completed, the route navigation is started, and the navigation function is also developed using the Gaode navigation SDK. Before starting the navigation component, you must first add the required permissions, then pass in the start point, end point and path point, and set the relevant parameters to navigate.

5. Conclusion Whether this angle
In this paper, data mining technology and collaborative filtering algorithm are applied to the travel recommendation system, which can provide users with current popular attraction ranking information, and can also recommend attractions to users according to the similarity between users. However, due to limited personal capacity, there are still some areas where further improvement is needed.

(1) Because it is not a professional origin, the interface design lacks some aesthetics, and the overall appearance is rather monotonous. In the future, it can be further optimized to make the interface look more beautiful and more in line with the user's aesthetic requirements.

(2) You can enrich the system functions, such as hotel reservations, attraction ticket purchases, location public toilet location inquiry, etc.

(3) Tourism is a relatively complex topic. This paper mainly studies based on the similarity between users. Later, it can be studied based on other angles, and then the recommendation results are analyzed and compared to obtain more accurate recommendation results.

References
[1] Liu, J.Zhou, T.Guo, Q.(2009)A Survey of Personalized Recommendation System Evaluation Methods [J], Complex systems and complexity science, 6(3):1-10.
[2] Xiang, L (2012) Recommended system practice [J]. People Post Press, Beijing.
[3] Wen, Y. (2014) A Summary of Personalized Tourism Recommendation Research [A]. In: 2015 China Tourism Science Annual Meeting. Beijing ,2-13.
[4] Liu, J.Fan, Y.(2014) The composition, value and development trend of smart tourism [J]. Chongqing Social Sciences, 16(10):121- 124.
[5] Zhang, M, Chen, Y. (2009) Study on the Recommendation Technology for Tourism Information Service[J]. International Symposium on Computational Intelligence and Design, 410-415.
[6] Shen, J.Deng, C.X.Gao, X.(2016) Attraction recommendation: Towards personalized tourism via collective intelligence[J]. Neurocomputing .173(3):789-798.
[7] H.H.Owaied, H.A.Farhan, N.Al-Hawamdeh, N.Al-Okialy.(2011)A Model for Intelligent Tourism Guide System[J].Journal of Applied Sciences.11(2):342-347.
[8] Xie, H. (2010) Tourism product recommendation system based on rough set [D]. Fudan University.
[9] Wu, L.Liu, L.(2006) A Survey of User Modeling Techniques for Personalized Recommendation Systems [J]. Journal of Information Science.25(l).2006(2).55-62.
[10] Yang,J. (2013) Research on User Interest Model and Real-time Personalized Recommendation Algorithm [D] Nanjing University of Posts and Telecommunications.
[11] Wen, J.He, B.Hu, Y.(2016) Research on Hybrid Recommendation Algorithm Based on Trust Degree of Social Network Users [J]. computer science, 43(1):255-258.
[12] Fu, Y, Qiu, Y. (2016) A Personalized Collaborative Filtering Recommendation Method Based on Bayesian Network [J]. computer science, 43(9):266-268.
[13] Cai, H, Chen, C, Wem, J. (2016). Research on Personalized Location Recommendation Algorithm Based on User Sign-in and Geographical Attributes[J]. computer science,43(12):163-167,178.
[14] Li, R.Li, M.Guo, W.(2016) Research on Collaborative Filtering Algorithm Based on Improved Similarity [J]. computer science ,43(12):206-208,240.
[15] Wan, J. (2013) Research and Improvement of Weighted Association Rules Mining Algorithm [J]. Guangxi University.