Research on Travel Personalized Recommendation based on Web Crawler

Haifei Ma, Lei Yang, Yilin Su, Yukun Li, Yungui Chen

Guangdong University of science and technology, Dongguan, Guangdong, 523000, China

Abstract

In view of the lack of 'special price' oriented function software in the current market, it is recommended to recommend air tickets, hotels and appropriate tourist cities, use Python language to compile web crawlers, crawl the air tickets, hotels and scenic spots of Ctrip and qunar, and use spring boot By analyzing the data and information obtained, the special price recommendation and popular scenic spot recommendation in the current season are realized, which fills in the vacancy of "economical and applicable" on large-scale tourism websites, and provides viewing direction for users who want to plan their own travel routes.

Keywords

Travel Recommendation System; Web Crawler; Data Analysis.

1. Introduction

Travelers can find travel information and share their travel experience through various channels and platforms. This information includes travel websites, social networking sites, blogs, forums, and various search engines, such as Google and Yahoo. A large number of recommendation algorithms[1-3] have been reported. Abdul Majid proposes a travel recommendation system based on personalized location based on personalized travel information provided by users[1]. Fenza G proposed a context-aware travel recommendation system based on collaborative filtering with fuzzy clustering[2]. Zhang Z K et.simulates the user's travel decision process based on the user's preference factors and geographic factors, and combines the influence of these two factors on a specific user to generate a recommendation list[3]. The above several travel recommendation systems and currently existing Web travel websites have not yet developed a "special price"-oriented functional module that recommends airline hotels and suitable tourist cities to users who are looking for economic and comfort. Currently using Python language to write web crawler[4] programs can successfully obtain big data on travel websites. In order to solve the contradictory demands of economy and comfort, travelers will make further decisions on "where to travel". This article selects Ctrip and Qunar as the data to be obtained, and uses the Spring boot framework to conduct a brief analysis of the obtained data and information, and implements the special recommendation of the season and the recommendation of popular attractions, which fills the large travel website. Affordable's vacancy, designed and implemented a special ticket hotel and recommended attractions system to meet the needs of users' personalized recommendations.

2. Design of Web Crawler

This article's crawler program is mainly divided into 2 parts: Design of URL and Data storage and processing respectively:

First, URL is the entrance of web crawler crawling webpage, and picking out a URL to be queried is very important for obtaining HTML source code of parsed page.

Data storage and processing: Web pages crawled by search engines through crawlers, connect to the database through MySQLdb, build tables and insert data into the original page database. Because data crawling is a time-consuming computer resource, before crawling data each time, it is necessary to determine whether the data to be crawled already exists in the database to reduce the number of queries and improve system performance. The crawler flow chart of this system is shown in Figure 1.

(1) Acquisition URL and web page information

Before crawling webpage information, you need to make sure that it is the webpage where the webpage is to be crawled, and find the required information for targeted crawling. In this article, the system selects the ticket information on Ctrip and the hotel information on Qunar as the data to be captured. After stitching the date/place of departure of the written url, the Google Chrome browser is used to analyze the webpage. Take the Ctrip URL as an example, use the Chrome browser to enter the ticket module in Ctrip.com, search for the departure city "Beijing" and arrive at the city"Shanghai", select "2022-06-03" on the date to search a total of 50 flight information, and pass After querying multiple dates, we can see that the URL changes are limited to: https://flights.ctrip.com/itinerary/ oneway/hfe-bjs?date=? The position of the question mark in the, this position represents the date to be queried, so you can define an expression to manage all the URLs to be crawled: https://flights. ctrip.com/itinerary/oneway/hfe-bjs?date= ' +str(date), use'str(date)' to match the input date, and set the date interval to control the crawling range when looping crawling.

(2) Whether data is retained in the database

Determine whether the table exists: Because the date of crawling data is selected as the table name, all table names ("show tables;") in the database are queried as a collection ([cur.fetchall()]) to determine whether the date exists If it does not exist, it means that it has not been created yet and data acquisition is required. Determine whether the data already exists in the table: Take the starting place as the selection condition (cur.execute()), execute the sql statement to determine whether the returned result is empty, if it is empty, it means that there is no data in the table, and data acquisition is required.

(3) Use Selenium to crawl pages

For a complex web page structure, key data on the page may only be displayed after user interaction, such as continuously scrolling down the scroll bar. For these data that are not easy to obtain, use the method of simulating a browser to capture, and Selenium is a browser automation tool that can directly operate the browser, and then get the source code after the web page is rendered. Using Selenium can solve more complicated web pages. Grab the task. Google Chrome 80.0.3987.149 combined with Selenium tool to achieve data crawling, in which the Google application (.exe) should be placed in the same folder of the Python program, if you want to modify the browser version used, you can modify it in the Python code, the crawl function Implementation includes obtaining source code and web page information analysis.
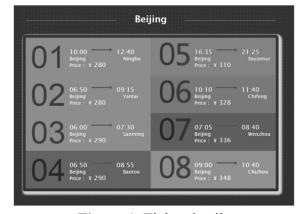
3. Experiment

Figure 1. Ticket details

Figure 2. Hotel details

The Recommadation system is based on querying and ranking airline tickets and hotels in different time, different flights, and different regions. as shown in illustration, Figure 1 Ticket details and Figure 2 Hotel detail.

4. Summary

This article analyzes the current large-scale tourism websites, and finds that there is a general lack of "special offer"-oriented, recommend air ticket hotels and suitable tourist city function module, based on this situation, designed and developed a travel recommendation system based on web crawlers. The system involves the analysis and processing of popular big data at home and abroad. A large amount of crawled data is analyzed and sorted according to the set conditions. It is concluded that the strategy of the tourist city (including ticket information, hotel information and Attractions information).

Acknowledgments

References

[1] A system for mining interesting tourist locations and travel sequences from public geo-tagged photos[J]. Data & Knowledge Engineering, 2015, 95:66-86.

[2] Fenza G , Fischetti E , Furno D , et al. A hybrid context aware system for tourist guidance based on collaborative filtering[C]// FUZZ-IEEE 2011, IEEE International Conference on Fuzzy Systems, Taipei, Taiwan, 27-30 June, 2011, Proceedings. IEEE, 2011.

[3] Cheng C C , Chen P L , Chiu F R , et al. Application of neural networks and Kano's method to content recommendation in web personalization[J]. Expert Systems with Applications An International Journal, 2009, 36(3p1):5310-5316.

[4] Pinkerton B , Lazowska E , Zahorjan J . Abstract WebCrawler: Finding What People Want[D]. Dissertation Abstracts International, Volume: 61-11, Section: B, page: 5967.;Chairpersons: Edward La, 2000.