Data Mining and Visual Analysis of Attractions

BoDan Ma

{bdm_000509@163.com}

Beijing Jiaotong University

Abstract. In this paper, the distribution of attractions and the number of attractions in each region of Beijing, Tianjin and Hebei province are counted and visualized by obtaining the names of all attractions and their detailed addresses in the Beijing Tourism Network. The analysis reveals that the distribution of attractions is not uniform and that Beijing has the highest number of attractions among the three.

Keywords: Data mining; visualization; distribution of attractions

1 Introduction

The spatial distribution pattern of regional tourist attractions and their multi-scale characteristics are an important basis for urban tourism development planning. In general, relevant scholars have formed a certain research system for the study of tourist attractions and tourism resources, but most of the existing studies focus on the region's important tourist attractions, lacking the exploration of the study of regional tourist attractions as a whole. When data mining methods are used to analyze tourist attractions, some scholars study the best tourist routes[1][2][6]; a scholar has conducted geospatial data mining based on text[4]; some scholars carry out personalized recommendation of tourist attractions[7-9]; some scholars analyze tourism data[3]; modeling analysis based on user profiles[5]; and research on tourists' behavior[10]. For visual analysis of the data, this paper uses Tableau software to make graphs to get a more intuitive sense of the distribution of the number of attractions.

This paper takes Beijing, Tianjin and Hebei tourist attractions as the research object, relying on the Beijing tourism website to obtain data, while using Seaborn and Tableau and other related tools to carry out visual analysis, to more intuitively feel the number of attractions and their distribution, in order to provide reference for tourists to make more reasonable travel plans.

2 Data acquisition

2.1 Data Acquisition in the Eastern District of Beijing

This data crawl decided to first try to crawl a particular district in Beijing, for example Dongcheng District. The steps are as follows:

(1) Import requests

“import requests

ICEKIM 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2337306

```
from lxml import etree”
```

(2) Enter Dongcheng District url

(3) Creating User-Agent dictionary and simulated browsers

```
headers = {"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) Apple WebKit/537.36 (KHTML, like Gecko) Chrome/94.0.4606.61 Safari/537.36 Edg/ 94.0.992.31"}
```

(4) Send request and return response

```
response = requests.get(url, headers=headers)
```

(5) Checking status codes of response, whether normal web data is received

```
if response.status_code != 200:
    print("Internet Error")
```

(6) Decoding

```
web = response.content.decode('utf-8')
```

(7) Converting web content into html formatted data

```
html = etree.HTML(web)
```

(8) Crawl for the name of the attractions

```
parse = ".//div[@class='info fl']/h5/a/text()"
html.xpath(parse)
```

Save data:exp_station1 = html.xpath(parse)

(9) Crawl site address

```
parse="//div[@class='adress']/text()"
html.xpath(parse)
```

Save data:exp_address_list1 = html.xpath(parse)

(10) Data collation

The data is collated as the information obtained is irrelevant and some addresses do not have "Beijing" in them, which would lead to errors in obtaining latitude and longitude.

(11) Check that the number of addresses matches the number of attractions

```
assert len(exp_address1)==len(exp_station1)
```

(12) Consolidation of all code into functional form

(13) Get the latitude and longitude of Dongcheng District attractions addresses

By combining all the code from the above steps into a functional form, you can obtain data about attractions in the Dongcheng District of Beijing.

2.2 Obtaining data on attractions in all districts of Beijing

(1) Get the name and address of the attraction

Direct input function, the function code is basically the same as obtaining the data of Dongcheng District, and the obtained data will be stored in a csv file; as the web page url is different for each district in Beijing, the number of attractions contained in each district is different and the number of page numbers is different, therefore, it is necessary to find the law to obtain uniformly.

For example, the url of the first page of attractions in Dongcheng District is "https://s.visitbeijing.com.cn/ attractions?area=1&page=1", the url of the first page of the remaining districts differs from that of Dongcheng District in terms of area, and the number of pages in each district does not exceed 50. The number of pages in each district does not exceed 50, so just changing the number in the code.

(2) Get latitude and longitude

The code is the same as getting the latitude and longitude of the East End attractions and storing the results in a csv file.

2.3 Obtaining data on attractions in Tianjin and Hebei Province

Collation of information on attractions replace "Beijing" with "Tianjin" or "Hebei", change url and area, and store the results in separate csv files “exp_attractionstianjin.csv” and “exp_attractionshebei.csv”; Store the results in separate csv files Store the address latitude and longitude results in separate csv files “exp_location2.csv” and “exp_location3.csv”.

3 Data visualization

3.1 Data visualization - Seaborn mapping

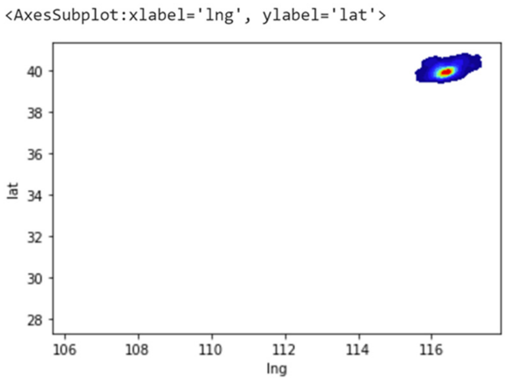
Visualization of the latitude and longitude of the acquired site addresses and density distribution maps using Seaborn.

(1) Visualization of addresses of attractions in Beijing

First get the color map, then pass the color map parameter to the kdeplot function.

```
df = pd.read_csv("exp_location1.csv")

cmap = plt.get_cmap("jet")

sns.kdeplot(x="lng",

            y="lat",data=df,fill=True,

            cmap=cmap)
```

Finally, the visualization result is shown in Fig. 1:

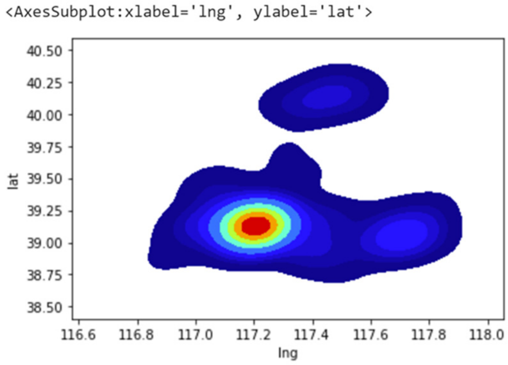


```
<AxesSubplot:xlabel='lng', ylabel='lat'>
```

Fig. 1. Seaborn mapping-Beijing

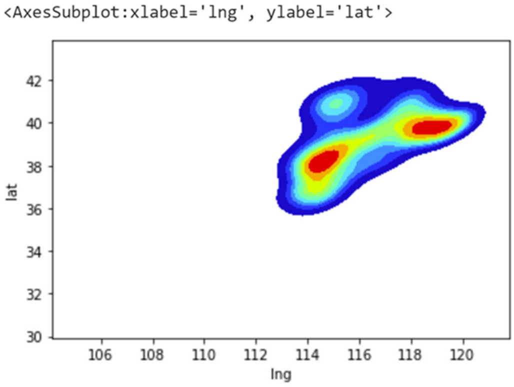
(2) Visualization of addresses of attractions in Tianjin and Hebei

The steps and codes for obtaining are the same as for Beijing, and the visualizations of the addresses of attractions in Tianjin and Hebei province are shown in Fig. 2 and Fig. 3:



```
<AxesSubplot:xlabel='lng', ylabel='lat'>
```

Fig. 2. Seaborn mapping-Tianjin



```
<AxesSubplot:xlabel='lng', ylabel='lat'>
```

Fig. 3. Seaborn mapping-Hebei

3.2 Data Visualization - Tableau Mapping

Based on the acquired csv files of the latitude and longitude of the 3 attractions, the distribution of the attractions was represented on the map using Tableau software. As shown in the Fig. 4, Fig.5 and Fig. 6:

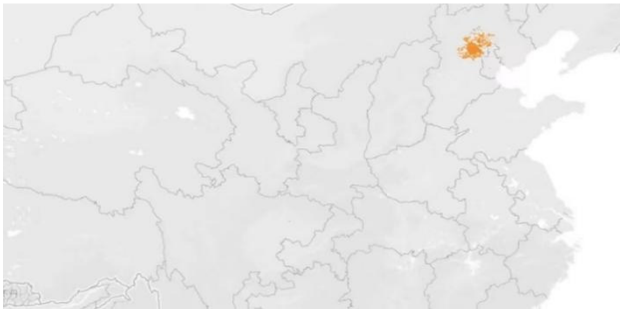


Fig. 4. Tableau Mapping-Beijing

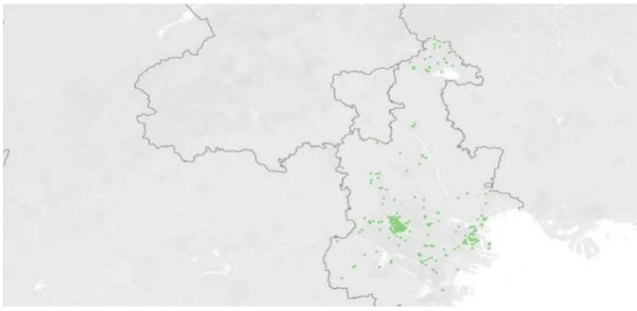


Fig. 5. Tableau Mapping-Tianjin

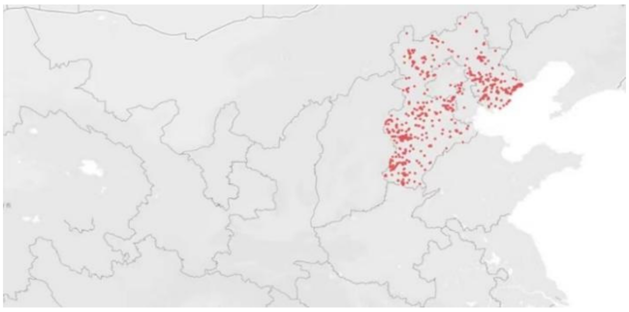


Fig. 6. Tableau Mapping-Hebei

4 Total number of sites acquired and visualized

4.1 Access to names and total number of sites by region

The total number of attractions in each of the 18 regions can be seen on the home page, so it is possible to crawl both the name of each region and the corresponding total number of attractions.

The code for crawling the names of the regions is: parse = ". //li[@class='active']/a/span/text()"; the code for crawling the total number of attractions in each region is: parse = ". //span[@class='total']/i/text()".

First, crawl the name of each region, the number of attractions, the name of the region, the number of attractions, then remove irrelevant information, call the csv file writing function, and finally, get the name of each region and its corresponding total number of attractions saved in the csv file.

4.2 Data visualization

Bar charts and histograms are plotted based on the data obtained, combining the two together, and the visualization results are shown in Fig. 7 and Fig. 8:

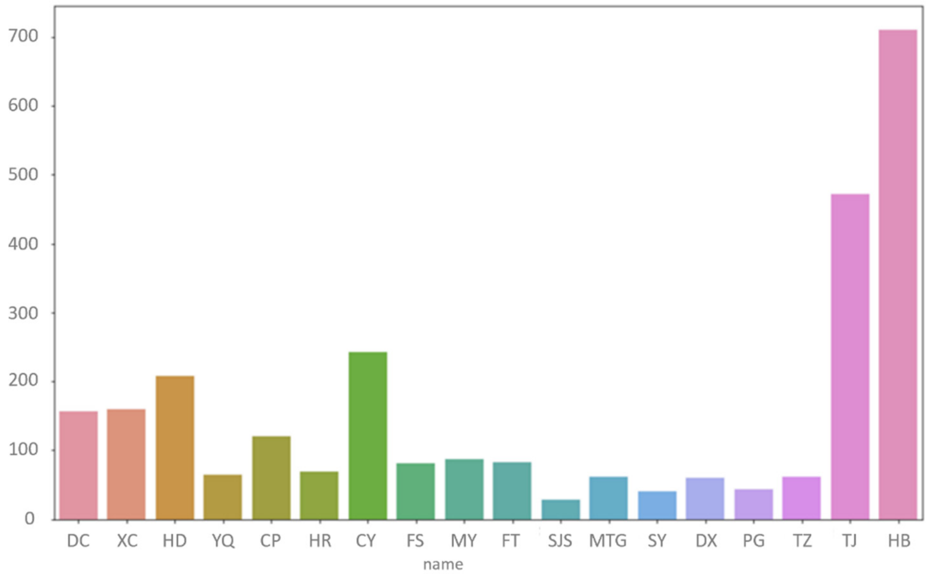


Fig. 7. Bar chart

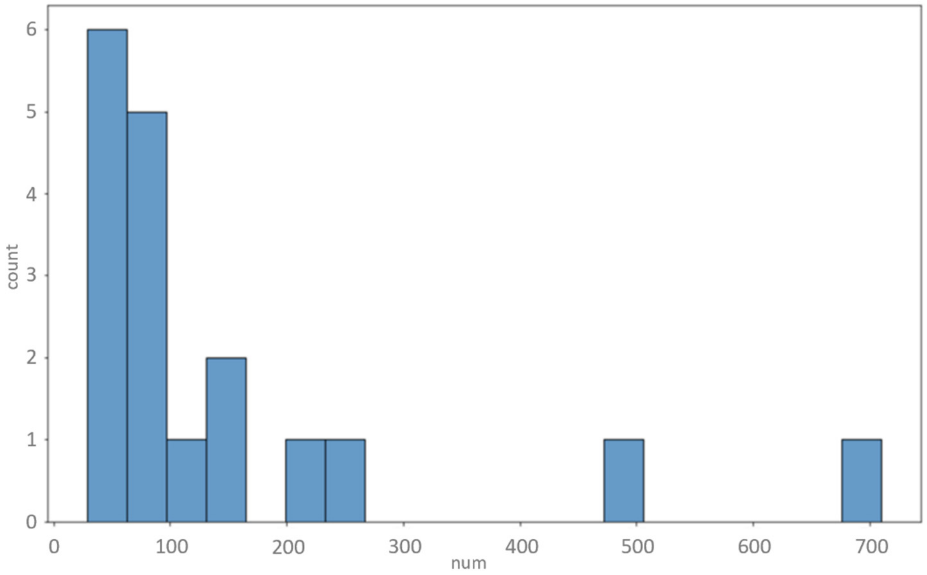


Fig. 8. Histogram

5 Conclusions

According to the obtained maps of the distribution of attractions in Beijing, Tianjin and Hebei, it can be seen that the distribution of attractions is not even, with some attractions concentrated in one area and others even absent.

Overall, Beijing, as the capital, has the highest total number of attractions in Beijing, Tianjin and Hebei. The histogram shows that the two rectangles on the right represent Tianjin and Hebei respectively. By looking at the six rectangles concentrated on the left, the total number of attractions in the 16 districts of Beijing occupies 11 out of less than 100, most of which are suburbs of Beijing, but among them, Changping District as a suburb has more than 100 attractions, while Fengtai District and Shijingshan District as the six districts of the city Chaoyang District and Haidian District are close to the city centre, each with a total of over 200 attractions, while Dongcheng District and Xicheng District are slightly behind Chaoyang District and Haidian District, both with a total of over 150 attractions.

Reference

[1] Xiao Z, et al. Tour Route Planning Algorithm Based on Precise Interested Tourist Sight Data Mining." IEEE Access PP.99(2020).DOI:10.1109/access.2020.3010420.

[2] Sun Tonghui. Research on the intelligent tourism route of Zhoushan based on data mining [D]. Zhejiang Ocean University,2019. DOI:10.27747/d.cnki.gzjhy.2019.000272.

[3] Gao Xinbo, Shen Junge. Social media-based tourism data mining and analysis[J]. Data Collection and Processing,2016,31(01):18-27.DOI:10.16337/j.1004-9037.2016.01.002.

[4] Liu Yuanfeng, Zhou Rongfu, Li Fengling. Text-based geospatial data mining and visualization[J]. Survey and Mapping Science,2010,35(04):103-105.
DOI:10.16251/j.cnki. 1009-2307.2010.04.043.

[5] Zhou Xiao, Zhou Xinghan, Yang Shuai et al. A spatial decision model for cultural tourism based on user portrait mining[J]. Green Technology,2022,24(24):276-280.
DOI:10.16663/j.cnki.lskj.2022.24.044.

[6] Gong Yanyuan. Design of the best tourism route planning system based on data mining[J]. Electronic Design Engineering,2020,28(09):59-62. DOI:10.14022/j.issn 1 6 74 – 6 2 3 6.2020.09.013.

[7] Pan Lusheng. Personalized recommendation-oriented tourism information mining for Gansu attractions [J]. Journal of Lanzhou College of Arts and Sciences (Nature Edition),2018,32(06):82-87.DOI:10.13804/j.cnki.20956991.2018.06.018.9.013.

[8] Zhang Zhanzhao. Analysis of the application of weighted mining algorithm in the recommendation system of intelligent tourist attractions[J]. Digital technology and application. 2017(03):168+256. DOI:10.19695/j.cnki.cn12-1369.2017.03.092.

[9] Ou Dan. Research and implementation of personalized travel recommendation system based on data mining [J]. Hubei Agricultural Science, 2021,60(09):123126.
doi:10.14088/j.cnki.issn0439-8114.2021.09.025.

[10] Wang Yun,Bai Yi. Research on tourists' online behavior supported by big data mining--Tianhetan scenic area as an example[J]. Journal of Guizhou Normal University,2019,35(12):23-27. DOI:10.13391/j.cnki.issn. 1674-7798.2019.12.006.