Geographical name culture sharing query service system for the "Belt and Road"

Meiao Wang [1], Wei Gao [2], Xinyuan Fu [3], Wenhao Sun [4, *]

[1] School of Management, Hefei University of Technology, Hefei, China

[2] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China

[3] School of Biological Science &Medical Engineering, Southeast University, Nanjing, China

[4] College of Civil Engineering, Hefei University of Technology, Hefei, China

* Corresponding author: 2019211239@mail.hfut.edu.cn

Abstract. With the advancement of the "One Belt and One Road" initiative and the rapid development of the world economy, given the problems of scattered geographical name culture resources and incomplete information in the areas along the "Belt and Road", this paper aims at the massive geographic information data along the "Belt and Road" and proposes a method based on web crawlers and task crowd. Package shared service platform construction method, and develop the client based on MySQL database and WebGIS, provide real-time data update and resolution, cultural information query and other functions, and realize geographic information and place name cultural query function. Open up new ideas for the construction of the "Belt and Road" platform, perceive and obtain geographical information of geographical names from the public's perspective, and provide real-time and dynamic inquiry channels for the public.

Keywords: Geographical name culture, "One Belt, One Road", Web crawler, System development.

1. Introduction

The "Belt and Road" is an important strategy to realize the development of international exchanges, which has special significance in promoting the construction of a community with a shared future for mankind.[1] Toponym information describes the names, locations, characteristics, and historical and cultural information of various geographic locations on the earth's surface, and plays an important role in navigation and positioning, social and economic development, and so on. The implementation of this policy makes the cultural and historical connotation carried by place name information more practical. Many scholars have begun to pay attention to the related fields of place names along the “Belt and Road”. For example, Zhou Ziyin integrated academic resources and artificial intelligence technology to build a platform to promote academic exchanges along the “Belt and Road” [2]. The Foreign Cooperation and Exchange Center of the Ministry of Ecology and Environment has built the "big data service platform for ecological and environmental protection" to realize the summary of various ecological and environmental protection documents [3]. However, the construction of these platforms focuses on the cooperation of enterprises, scientific research institutions, and government departments, and the general public has a relatively low sense of participation. In fact, encouraging, supporting and guiding the public to participate in the "Belt and Road" is an important measure to enhance the influence of the "Belt and Road".

Therefore, the establishment of a geographical name culture sharing service system for the “Belt and Road” can help users in countries along the route to share geographical name cultural information, to better understand each other’s history, culture, and geographical environment [4]. This will help promote cultural exchanges and understanding, and provide a better basis for economic and cultural communication among countries along the route. At the same time, using the form of crowdsourcing of data tasks, uploading and modifying data by users, based on massive non-specialized spatiotemporal data sensors such as smartphones and cloud computing technology, greatly enriches

spatiotemporal data information resources, forming a world where everyone is a geographic information data sensor A new situation [5].

Based on this, to enhance the current situation and pertinence of the service platform, and enhance the practical significance of geographical information display of place names, it is convenient for the general public to understand the "Belt and Road", enter the "Belt and Road", and benefit from the "Belt and Road"[6]. We use the multi-source data acquisition method based on web crawlers and task crowdsourcing, and regularly update technology to build a shared service platform for the public.

2. Data Processing

2.1. Data Sources

The basic data for the construction of the shared service platform comes from two major carriers - website and APP. For websites, we mainly focus on the GeoNames website [7], which collects more than 10 million place names, including 2.8 million well-known places and 5.5 million aliases. In addition, based on Google search engine services, using web resources as data sources, using web crawler technology to obtain a large amount of relevant data, and then through place name recognition and spatial location analysis technology, the maintenance and update of geographical name database can be realized. It can also use the data of existing scientific global geodatabases (GeoNames), Geographical Names Network Service (GNS), Geographical Names Information System (GNIS), Location Database of Administrative Regions of the World (GADM), Geographical Names Database (GADM) of Alexandria, Geographical Names Database (GOA) of Australia, World Gazetteer in the 21st Century and other authoritative institutions. In addition to the above data sources, we also make full use of volunteer geographic information platforms such as Wikipedia, Open Block Map (OSM), and the World Administrative Divisions Network to supplement the data of countries and regions along the "Belt and Road". Using crawlers, we focus on collecting domestic open-source geographic information platforms, to enrich the geographic information database. For the APP, we mainly chose the "Meituan" APP, which highly integrates a number of life information including restaurants, pharmacies, various shops, and tourist attractions, and has geographical attributes due to its mobile GIS characteristics.

The data from the two major carriers of websites and APP has great characteristics. Take GeoNames as an example: In the main table of geoname, it includes information such as region code, name, latitude and longitude, alias, zip code, and time zone, and the data breadth is large. However, it is not difficult to find that these are the most basic geographic information data, which are not highly relevant to life, not close enough to culture, and have poor timeliness. For example, "Dawa County" in Panjin City, Liaoning Province, was renamed "Dawa District" in 2016, but it is not updated in the database. In addition, due to language differences, when faced with the geographical names of the autonomous region such as "Horqin Right Wing Front Banner" and "Horqin Left Wing Middle Banner", it is impossible to subdivide them, which increases the roughness of the data. At the same time, Internet companies such as Gaode, Meituan, and other r Internet companies have also collected rich geographic data information within their business scope. Taking the "Meituan" APP as an example, the takeaway, travel and other services it covers involve many geographical information such as catering stores, which has high use value, and also contains massive merchant information and its evaluation, with large data depth. However, "Meituan" is consumption-oriented, insufficiently oriented to geographic information, focusing on product promotion, and a large number of advertisements will reduce the user experience, and can only be used for domestic users, and the scope is limited. Therefore, the integration of the two will further improve the quality of the service platform in the two dimensions of geographic information and culture, and display richer and more delicate information for users.

2.2. Data Acquisition

With the development of information technology, countries in the world have successively established their own and even global geographic information databases, and the development of big data has accelerated this process. This paper uses the online acquisition of public datasets and web crawlers as the main data acquisition methods [8]. In addition, data methods such as semi-structured text parsing and vector file parsing are proposed as supplements. Well-known government agencies, businesses, and organizations choose to open source their data so that other developers can use their data directly. For example, Natural Earth, GeoNames and other datasets.

However, not all geographic information data is displayed in the form of datasets or APIs, so there is still a lot of geographic information data to be obtained. Collecting the required toponymic data from the vast amount of information obviously costs a lot of manpower and material resources, and the development of the Internet has provided a solution to this problem [9]. As a computer program (or automated script), Web crawler can obtain, process, and store data on the Internet according to the rules set by the developer, and is an important part of the search engine. The categories of web crawlers mainly include general web crawlers, focused web crawlers, incremental web crawlers, deep web crawlers, etc., through depth first, breadth-first or best priority algorithms for web page information retrieval, currently widely used in search engines, advertising filtering, big data analysis and other fields [10].

The paper encodes web crawlers for different web pages and retrieves place name information through this. The main processes include:

(1) Select the target of the page to be crawled as a seed URL;

(2) Retrieve all URLs in the main page, such as <a>hyperlinks in HTML as tags;

(3) Repeat step 2 to iteratively crawl the URL, for example, you can extract the key information on the page until you reach the deepest geographic information display unit. The information that can be crawled includes introductions, latitude and longitude, zoning codes, etc., as well as unique data that may be available on certain pages;

(4) After doing a good job of web page analysis, data collection and recording, go back to the previous level and repeat the process. Figure 1 depicts the main process of information acquisition tasks performed by web crawlers:

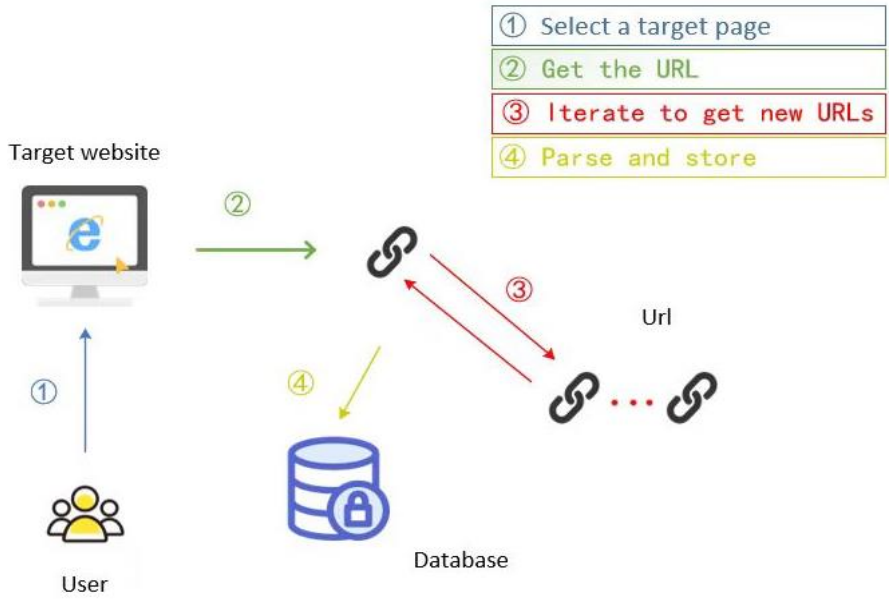

Figure 1. Schematic diagram of using web crawler to collect place name information

Among them, it should be noted that because the data presentation of different web pages may be different, the specific structure of the crawler will be different. When the data is obtained by the web crawler, the data is checked again to review possible vulnerabilities and errors, such as empty data, malformed data, and so on, programmatically or manually.

Considering the many different data sets we use and the crawling situation of web crawlers, there is bound to be duplication of data in different data sets, and this simple data listing is not organized in any way, so data cleaning is required. We consider the following situations and take appropriate action respectively: First, identical or equivalent data items. When the data between different datasets is completely consistent or consistent after conversion due to language, date format and other issues, redundant items are removed; Second, conflicting data will be stored in specific data tables for review and subsequent errata; Third, complementary data information will be stored in the "Other" item in the corresponding table.

2.3. Data Meditation

With the increasing speed of data collection and the increasing amount of data, to ensitize the toponymic database to make full use.

First, read the data uploaded by the user, obtain the metadata, start the sensitive data identification engine, and use the extraction of feature words, regular expression matching, feature density calculation, etc. set in the background in advance to intelligently identify sensitive data, and store the identified sensitive data in a specific database. Based on the sensitive field database obtained in the previous step, reconfirm the desensitized data object, that is, sort out the basic information including business objects, basic tables, field names, and association relationships between tables. Create different user permission levels, to set up hierarchical desensitization functions in the data desensitization system to establish different levels of desensitization models, desensitization models include but are not limited to desensitization process framework, desensitization attributes, desensitization function rule designation and desensitization algorithms, abnormal data processing methods, etc., database administrators can dynamically update more adaptable desensitization models according to different needs, and finally enable users with different permission levels to obtain information with different sensitivity, with flexible data granularity and control strategies. Reliable protection of sensitive and private data.

For example, when users upload latitude and longitude data, sensitive personal location information may be transmitted. At this time, the background regular expression is used to identify the data for the matching of latitude and longitude, store it in the sensitive database, process the latitude and longitude data in a truncated way, and store it in the public database after processing, to ensure the security of user positioning information.

3. Database Designing

The information stored in the geographic information database includes geographical name information, cultural information, thematic information, etc. Developed based on the Windows operating system and open-source MySQL database. The basic geographical name information in the database includes geographical name, area, latitude and longitude, population number, former name, division code, population density, etc.; As our featured function, cultural information is cleansed and integrated based on the data of the open-source platform, mainly collecting specialties and food, famous tourist attractions, related celebrities, related honors, etc. In addition, for administrative units at the city and district levels, the corresponding population rankings of each county and district, the ranking of the population of counties and municipalities nationwide, and so on are provided. Users can perform data errata, upload their basic information such as place names, areas, and cultural information to supplement and update, submit them in the form of text, pictures, etc., and will be updated in the database after verification by the administrator. The logical structure of the database is shown in Figure 2.

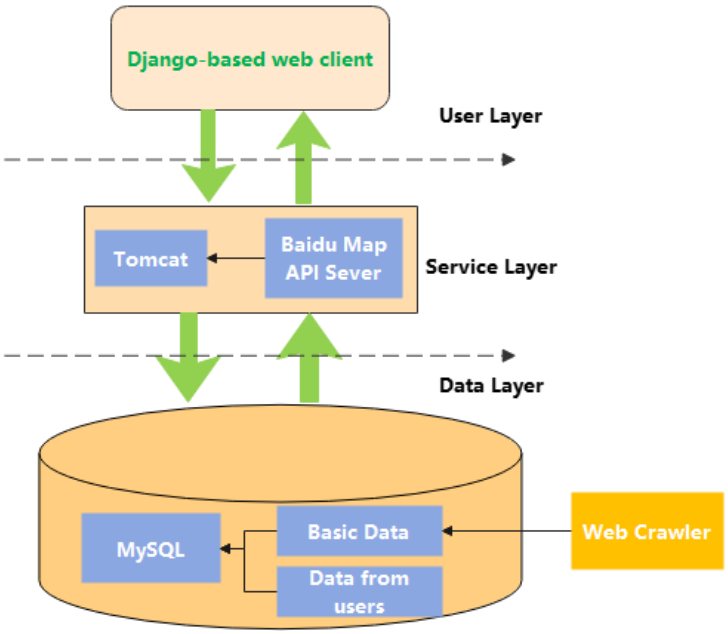

Figure 2. System Architecture

To facilitate the management of user information, the user's personal information is mainly stored in the user table, including user name, password, application update content, personal data, reward points and other data. In order to facilitate discussion and evaluation, user comment information, administrator response information, discussion time, processor ID, and events are provided in the communication form.

4. Construction of the platform

4.1. Data Meditation

In this paper, the shared services platform is designed based on the B/S architecture. Through the mode of data layer-> service layer-> user layer, data storage is carried out, and basic functions such as database addition, deletion, modification and query required by users are realized, and on this basis, multi-dimensional information display and data update are realized, with strong adaptability.

The basic geographical name is geographic information data, user information, and takeaway data of the data layer. Through the open-source MySQL database, all kinds of information storage and management are realized, and the optimal storage of user-updated data is realized. Connect databases to services with the help of JDBC tools in Java.

The service layer consists of an Apache Tomcat server and Baidu map server based on Docker container, using IntelliJ IDEA combined with JavaJDK platform to develop local servers, and Baidu Map API completes the map operation part.

The client layer consists of a client and a web client, relying primarily on the Django and Chrome platforms. Among them, the HTML development language and JavaScript are embedded in the Baidu Map API, and the Echart component is used to optimize the information display, and the user operation is communicated with the server through the HTTP protocol.

4.2. System function implementation

The website client of this article system is based on the Django framework, developed with IntelliJ IDEA through Visual Studio Code, and realizes map operations by calling the Baidu Map interface, communicates with the server based on HTTP, and the server uses the Apache Tomcat platform based on the Docker container to develop, and the database is organized and built using MySQL.

Figure 3 shows the login screen of the system.

Figure 3. Login System

(1) Overview of basic geographic information

As shown in Figure 4, the system mainly visualizes the cultural tourism scenery, special product overview, and geographical name changes of the geographical name cultural division. Users can freely choose major sections to view, query geographical name data, customs and culture, etc., and can also check relevant information in the "Belt and Road Column".

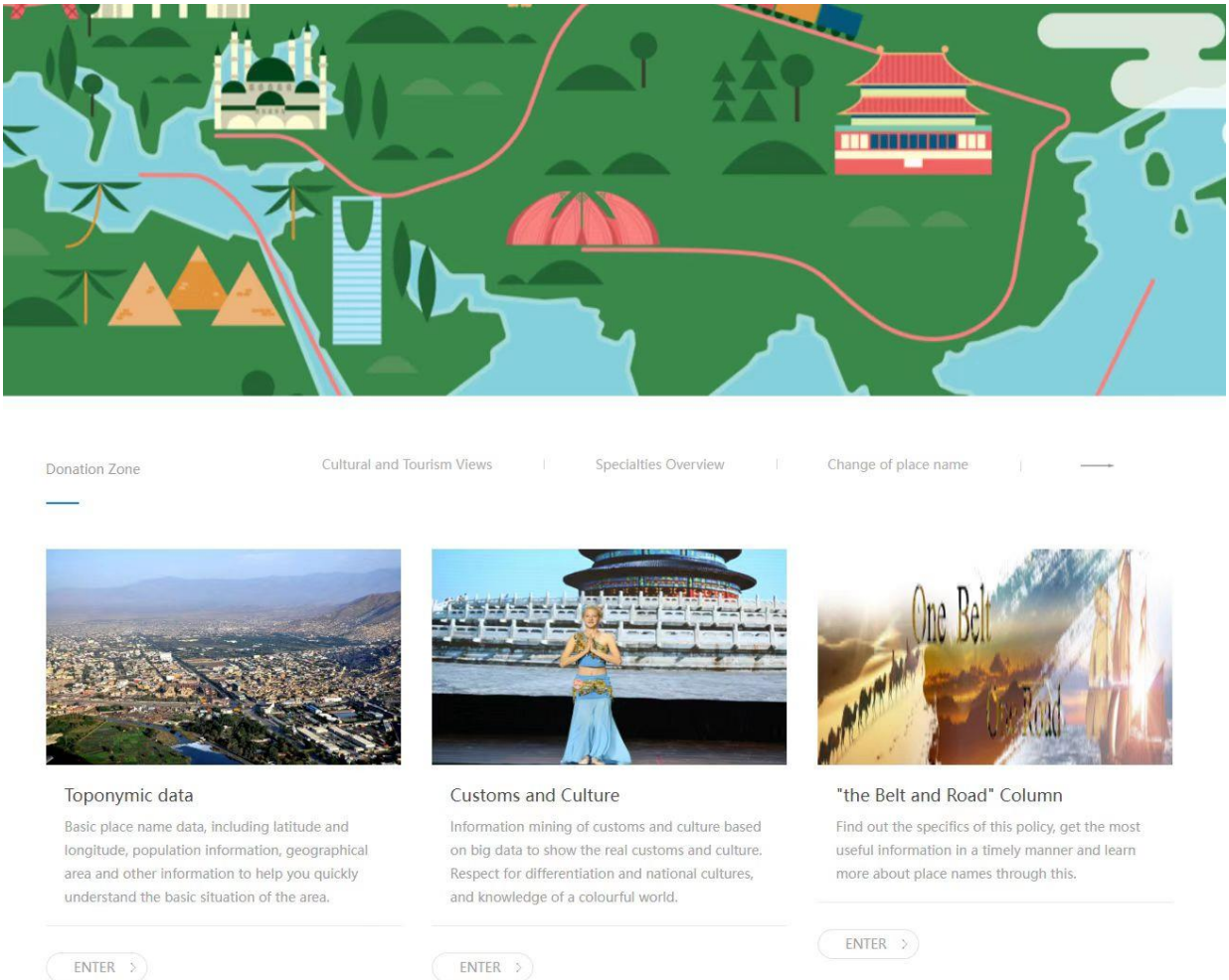

Figure 4. Overview of basic geographic information

(2) Overview of cultural tourism scenery information

The scenery of cultural tourism can better reflect the local spiritual outlook, and it is also one of the focuses of the general public in the face of unfamiliar areas. For example, Figure 5 shows a tourist attraction in Pakistan, a country along the Belt and Road, which users can click to view.

Figure 5. Overview of cultural tourism scenery information

(3) Overview of Specialties

The abundance of specialties reflects the customs and lifestyle of a region. Since our database contains some of the merchant and review information of food delivery platforms, it can be used as a reference for local product information. This system supports some special products with high ratings and high popularity in relevant takeaway platforms, and displays corresponding pictures and introduction information, aiming to provide users with more comprehensive and convenient cultural tourism services.

(4) Data Management and Update

Users are divided into ordinary users and managers, of which ordinary users are the main users who use the system, have low permissions, and can query and add basic data such as place names, latitude and longitude, pictures, etc.; The manager is a professional with high permissions, and can modify and delete data on the basis of ordinary users, and review the data uploaded by ordinary users, and only after confirming that it is correct can it join the public database.

5. Conclusion

Comprehensively use geographic information system, mobile Internet, database development and other technologies to build a basic geographical name database based on the background of the information age and the promotion of the "Belt and Road" policy, develop a visual information query website based on the web terminal, and build a multi-source geographical name database and shared service platform for the "Belt and Road". In order to guide more people to understand the "Belt and Road", participate in the "Belt and Road", and benefit from the "Belt and Road", form new ideas for the development of geographic information application, and promote the socialization, popularization and intelligent development of surveying and mapping geographic information science. With the rapid development of big data and artificial intelligence, in the future, related technologies can be used to achieve more fine-grained collection and screening of geographical name information, so as to better serve the "Belt and Road" policy.

Acknowledgement

The authors gratefully acknowledge the funding by the National College Students Innovation and Entrepreneurship Training Program (202110359037).

References

[1] Huang Y. Understanding China's Belt & Road initiative: motivation, framework and assessment [J]. China Economic Review, 2016, 40: 314 - 321.

[2] Zhou Ziyin. Research on the Construction of Big Data Management Service System of the New Think Tank Platform of "The Belt and Road" [D]. Shaanxi Normal University, 2018.

[3] Ministry of Ecology and Environment. Big Data Service Platform for Ecology and Environmental Protection.

[4] The Routledge Handbook of the Belt and Road [J]. 2022.

[5] Aripov N, Kamaletdinov S, Tokhirov N, et al. Simulation Modeling of Train Traffic Based on GIS Technologies [C]//AIP Conference Proceedings. AIP Publishing LLC, 2023, 2612 (1): 060025.

[6] Kumar N, Gupta M, Sharma D, et al. Technical Job Recommendation System Using APIs and Web Crawling [J]. Computational Intelligence and Neuroscience, 2022.

[7] GEONames: http://www.geonames.org/.

[8] ResearchGate. Geographic information systems: applications and research opportunities for information systems researchers [EB/OL]. 2022.

[9] Michler J D, Josephson A, Kilic T, et al. Privacy protection, measurement error, and the integration of remote sensing and socioeconomic survey data [J]. Journal of Development Economics, 2022, 158: 102927.

[10] Zhang, Y., & Wang, Y. (2019). Design and Implementation of a Query Service System for Shared Geographical Names Based on “One Belt One Road” [J]. Journal of Computer Applications, 39 (1), 1 - 6.