International Journal of Science and Engineering Applications
Volume 15-Issue 05, 80 – 81, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1505.1011

Design and Implementation of Postgraduate Examination Prediction and Recommendation Based on Python

Chaoxuan Chen
School of Geosciences,Yangtze University
Wuhan, China

Abstract: This system is based on big data technology, integrating Hadoop and Spark for data processing, Python for crawling postgraduate entrance examination information, Spring Boot and MyBatis for building the backend, Echarts for visualization, and MySQL for data storage. By analyzing application trends and the development of professional master's programs, it constructs a postgraduate entrance examination score prediction and school selection recommendation system to provide accurate decision support for candidates.

Keywords: Postgraduate entrance examination data analysis; big data development; Python development

1. INTRODUCTION

Since the 1990s, the enrollment of graduate students in Japan has seen a significant increase, rising from 90,238 in 1990 to 153,423 in 2010. This reflects the promotion of demand for higher education by economic development, improvement in material living standards, and the popularization of the concept of lifelong learning. However, the proportion of graduate students in Japan's total population remains relatively low compared to other developed countries. In contrast, China's graduate education started relatively late, and there is still a gap in its overall development level. Against the backdrop of accelerated global informatization and increasingly fierce competition, society's recognition of knowledge updating and the value of higher education continues to deepen, driving the continuous improvement and expansion of graduate education.

2. REQUIREMENT ANALYSIS AND SYSTEM ARCHITECTURE DESIGN

This session mainly focuses on the design of the overall project functionality, with the design plan primarily consisting of a big data system and a visual system comprising both front-end and back-end web management. In the web management system, the Springboot framework and Mybatis framework are primarily utilized. In the visualization phase, Echarts is employed to provide interactive and intuitive data visualization charts. The system utilizes MySQL as the database, which is designed to store a vast dataset of postgraduate entrance examination information crawled by a crawler, as well as the analysis results obtained after data processing. Data analysis is accomplished through Spark parallel computing, which facilitates operations such as data extraction, multidimensional analysis, query statistics, and more. The system's data detail query function retrieves the data analysis results from the MySQL database and ultimately generates Echarts charts to be displayed to users.

3. Detailed design and implementation of the system

3.1 Data acquisition function design

The data collection for postgraduate entrance examination utilizes the Scrapy crawler framework. By analyzing URL patterns, it generates target links and stores them in a dictionary. The original HTML data crawled is consistent with that of a browser, and care must be taken to avoid triggering anti-crawling mechanisms due to duplicate content. Subsequently, data cleaning is performed to remove redundant tags and invalid information, enhancing data usability and cleanliness, and providing a high-quality data foundation for subsequent analysis.

3.2 Data processing and analysis function design

When processing the postgraduate entrance examination dataset based on the Spark framework, first, create a Spark Context and connect to the resource manager to start the executor; then distribute the application to the executor, construct a DAG graph and divide it into multiple stages, with tasks being distributed by the task scheduler; after the executor completes the computation, it releases the resources, achieving efficient distributed data processing.

3.3 Distributed database

This system employs a distributed database to support efficient storage and flexible expansion of postgraduate entrance examination data. Compared to centralized databases, it can dynamically add or remove nodes based on changes in data structure without requiring a complete architecture reconstruction. The distributed database consists of multiple independent databases interconnected via a network, enabling transparent access. Its core technologies include data sharding (horizontal sharding splits data by rows, while vertical sharding splits data by columns) and a data synchronization mechanism. The former enhances storage and query efficiency, while the latter ensures data consistency across multiple nodes, thereby effectively supporting reliable management and analysis of postgraduate entrance examination information in a big data environment.

International Journal of Science and Engineering Applications
Volume 15-Issue 05, 80 – 81, 2026, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1505.1011

3.4 Design of the school recommendation function for postgraduate entrance examination

Collaborative filtering is divided into two categories: user-based collaborative filtering (UBCF) for finding similar users for recommendation, and item-based collaborative filtering (IBCF) for calculating school similarity for recommendation. Each has its own advantages and disadvantages, and is suitable for the scenario of choosing graduate schools for entrance exams.

3.5 Design of the function for predicting the number of candidates for postgraduate entrance examination

To maximize the accuracy of predictions, the KNN prediction algorithm was employed. KNN is a non-parametric and lazy algorithm that does not require assumptions about data distribution. It is suitable for both classification and regression, and its principle is simple and intuitive. However, it has high memory consumption, slow prediction speed, and is extremely sensitive to noise and outliers, which can easily lead to inaccurate predictions.

4. Detailed implementation of the system

4.1 Data import implementation

When it comes to manipulating data in a MySQL database in Python, it is necessary to install a library called PyMySQL. After installation, use the command `import pymysql` to import the Pymysql library. Then, call `Pymysql.Connect()` to obtain the connection object `conn`.

4.2 Data analysis using the Spark framework

This section primarily focuses on utilizing the Spark framework to perform parallel computing, multidimensional analysis, decision charts, and other operations on the postgraduate entrance examination information dataset that has been stored in Hadoop's HDFS. Before conducting a detailed analysis of the data, we first need to use Spark Session to read the data in preparation for the subsequent detailed data analysis.

4.3 Realization of big data visualization for postgraduate entrance examination

Visualize data with Echarts, an open-source visualization library implemented in JavaScript. ZRender runs smoothly on both PCs and mobile devices, is compatible with most modern browsers (IE8/9/10/11, Chrome, Firefox, Safari, etc.), and utilizes the default vector graphics library to provide intuitive, interactive, and highly customizable data visualization charts. As shown in Figure 1 below.

Figure 1. Echarts visual data chart

5. CONCLUSION

This system is designed for graduate students preparing for entrance exams. It utilizes collaborative filtering (including User-Based Collaborative Filtering, UBCF, and Item-Based Collaborative Filtering, IBCF) and the K-Nearest Neighbors (KNN) algorithm to provide intelligent school recommendation. Collaborative filtering leverages user or school similarity to mine preferences, while KNN predicts targets through neighboring samples. The combination of the two takes into account both personalization and accuracy. The advantage of this system is that it does not require strong data assumptions and can adapt to multiple scenarios. However, it faces challenges such as high computational cost and sensitivity to data quality. The overall design meets the actual needs of school selection.

6. REFERENCES

[1] Tian Xiao. Research on Computer Application Technology in the Big Data Environment [J]. Computer Knowledge and Technology, 2019 (14): 246-247

[2] Xu Chengjie. Design and Application of a Big Data Analysis Platform Based on Spark [M]. Journal of Health Information Management, 2019.16(05)

[3] Cao Haiping. Analysis of Big Data Mining Technology Based on the Spark Platform [D]. 2022, 43(07)

[4] Zhou Zhengyu. Design and Implementation of a Data Analysis and Visualization Platform Based on Spark [J]. Computer Knowledge and Technology, 2022, 18(24)

[5] Song Guoxing. Research on Improved Parallel K-means Algorithm Based on Spark Streaming [J]. Modern Computer, 2021(18)

[6] Zhang Ruoyu. Python Analysis of Scientific Computing [M]. Beijing: Tsinghua University Press, 2016

[7] Du Penghui. Design and Implementation of a Web Crawler Based on Scrapy [J]. Electronic Design Engineering, 2019.27(22)

[8] Li Jing, Wu Ziyi, Wu Yubao. Visual Analysis of Enrollment Data Based on Echarts [J]. Popular Standardization, 2022(16)

[9] Zheng Jiming, Liu Qing. Application of Echarts in Data Visualization Courses [J]. Computer Knowledge and Technology, 2020(02)

[10] Hou Congcong. Application of Computer Software Technology in the Era of Big Data [J]. Computer Knowledge and Technology, 2018(14)