Design and Application of Distributed University Books and Archives Management Information System Based on Hadoop

Ping Wang(✉)

Shandong Institute of Commerce and Technology, Jinan 250100, Shandong, China
798441577@qq.com

Abstract. Under the background of building a "digital campus" in an all-round way, the application value of books and archives management in colleges and universities is getting higher and higher, and the information transformation of related services and management has become an inevitable trend. In this regard, this paper takes the management of books and archives in colleges and universities as the research object, and puts forward a set of construction scheme of distributed management information system based on many problems faced at present, so as to better collect, protect, develop and utilize the information of books and archives. The platform uses Hadoop cluster to complete the distributed storage and management of books and archives data information, cooperates with Lucene algorithm tools to complete the design and development of retrieval engine, and integrates javaweb technology to form a standard network application program. The construction of the system can fully realize the digitalization of university books and archives resources, the networking of information services and the standardization of electronic archives management. The simulation test results show that the system effectively improves the utilization efficiency of books and archives resources, optimizes the information service process, and makes a positive contribution to the construction and development of colleges and universities in the new period.

Keywords: big data technology · books and archives management · Hadoop · Lucene algorithm · computer software application

1 Introduction

Under the background of building a Chinese-style modern education system in an all-round way, high-quality development has become the core task of deepening education reform. [1] In the new era, the management of books and archives in colleges and universities, as an important part of the management system in colleges and universities, can take the university library as the carrier. On the one hand, it can provide high-quality education, scientific research or education and teaching services for teachers and students, on the other hand, it can complete the original records of school history and school culture from a global perspective. [2] However, the traditional management

P. Qi and Z. Chen (Eds.): ICBDIE 2023, AISR 178, pp. 426–430, 2024.
https://doi.org/10.2991/978-94-6463-238-5_60

mode of books and archives can not meet the increasingly high-speed and frequent sharing and application of information resources. The problems of insufficient overall collection, low retrieval efficiency and poor disaster tolerance have seriously affected the management of books and archives in colleges and universities, so that the quality of books and archives management is low. In view of this, this paper believes that colleges and universities should pay more attention to the management of books and archives, persist in innovation-driven, and actively adopt the practical advantages of information technology such as the Internet, big data and databases. By building a distributed management information system, colleges and universities can promote the network and digital transformation and upgrading of the traditional management mode, and improve the overall level of books and archives management in colleges and universities. [3] The system takes big data technology as the core, integrates javaweb technology to build a server system supporting remote access, provides convenient data information services for users of different roles, and makes a beneficial attempt to realize the construction of "digital campus" as soon as possible.

2 Development Process

The overall development of distributed university books and archives management information system can be divided into two parts. One is to complete the construction of distributed data processing server with Hadoop framework as the core, and complete the deployment of Lucene data retrieval engine to meet the design requirements of system database layer. Hadoop framework will adopt cluster deployment according to data volume and functional requirements. [4] First of all, Hadoop cluster contains three functional nodes, namely Master1, Slave1 and Slave2, and each node needs a computer equipped with an 8-core hexadecimal CPU, 16G memory and 1TB hard disk as a server. Secondly, the server of each node adopts Linux Ubuntu 16.04 operating system, the JDK chooses dk-8u162-linux-x64, and the Hadoop framework version is v2.6.1. After Hadoop is installed, the components such as FileSystem, Replicas, Mapreduce, Yarn and HDFS are set and adjusted in turn to complete the construction of Hadoop environment. [5] Finally, the Lucene 3.0.3 tool is integrated with the Hadoop framework, and Mapper and Reducer are used to complete the operation control of the retrieval algorithm, so as to realize the construction of indexes by IndexWriter and IndexReader, and improve the retrieval efficiency of the system for massive book archives data information. [6] The second is to complete the encapsulation of system display layer, business logic layer, data access layer and database layer with javaweb technology as the core, and form a standard network application program. [7] Java is selected as the basic development environment, MyEclipse V 2022 as the integration tool, Tomcat 8.0 as the Web server and MySQL 5.7 as the database server. Through the introduction of the above key technical theories, the overall environment of system development, the configuration of related software and tools are determined, and the technical feasibility of the overall project of distributed university books and archives management information system is also clarified.

3 Functional Implementation

3.1 User Side

According to the actual functional requirements, the system presupposes three roles of users: student, teacher and administrator. After the user logs in successfully, the system will automatically jump to the homepage interface and complete the function navigation. The functional models of students and teachers are similar, mainly focusing on the retrieval, viewing, collection, downloading and other operations of books and archives data information resources.

When a student or teacher user inputs a keyword bar in the retrieval interface, the platform will segment the input content and calculate the similarity weight of the keyword bar in the index database by using TF-IDF formula. [8] Formula 1 is the formula for calculating TF-IDF, where TF represents the frequency of this keyword bar in an index, Nω is the number of keyword bars, and N is the total number of entries in an index. IDF represents the frequency of keyword bars appearing in all indexes, Yω is the number of indexes containing keyword entries, Y is the total number of indexes in the index database, and λ represents the similarity weight. The whole operation process depends on the IndexSearcr method in Lucene framework. The calculation results show that the greater the weight, the higher the matching degree between the content in the index library and the keyword bar. [9] According to the matching degree between the index and the keyword bar, the Web server will retrieve the corresponding book archive data information and return it to the client page for display, that is, return the user search results.

$$\lambda = TF * IDF \; TF = \frac{N_\omega}{N} \; IDF = \log \frac{Y}{Y_\omega + 1} \tag{1}$$

3.2 Administrator Side

Administrator users can convert paper books, newspapers, magazines, files, photos and other contents into digital resources and complete the uploading operation. At the same time, the existing CD-ROMs, electronic documents and other contents are also filed for processing, and the classified management mechanism of books and archives is improved [10].

Under the data statistics function, administrators can quickly grasp the retrieval details of books and archives resources by teachers or student users in a certain period of time, which provides convenience for exploring users' needs, realizing information linkage and promoting the level of information service. Figure 1 shows the recent keyword search rankings.

After the platform design is completed, various functions of the system will be simulated and tested. 356711 groups of data were tested, and Lucene framework was selected to compare with traditional SQL commands in retrieval efficiency. The experimental results are shown in Table 1. The results show that the search efficiency of Lucene framework is basically on the millisecond level, far exceeding the traditional

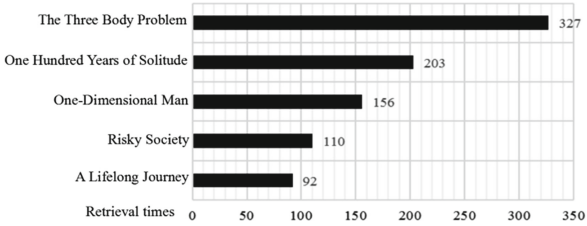

Fig. 1. Ranking of the key entries retrieved

Table 1. Comparison test results of system retrieval efficiency

Field name	Key entry	SQL command time (ms)	Lucene frame time (ms)	Judge
ID	157%	157	33	Consistent
Name	Li Li%	8300	1312	
Msg	Digital China%	10798	1524	

Table 2. Platform performance test results

Initiation time	Thread	Label	Connect time (ms)	Delay time (ms)
2022.3.13.10:53:03	1–1	Network request	11	35
2022.3.14.11:28:21	1–2	Network request	15	41
2022.3.14.12:13:59	1–3	Network request	13	66

SQL commands. In addition, the running performance of the system under high concurrent requests will be tested. In the experiment, 500 concurrent network requirements are simulated, and the network processing delay results of the system are shown in Table 2. The test results show that the system runs smoothly, the network delay time is controlled below 100ms, and the performance meets the design specifications.

4 Conclusions

In order to promote the reform of books and archives management system in colleges and universities, this paper aims at many shortcomings under the traditional management mode, and builds a distributed books and archives management information system in colleges and universities with the help of the practical characteristics of big data technology, network information technology and computer application technology. The system fully realizes the digitalization of university books and archives resources, the networking of information services and the standardization of electronic archives management. In the follow-up research, the actual interactive function of the system will be further expanded, and the application scope will be expanded, making positive contributions to the construction and development of colleges and universities in the new period.

References

1. Li Shulin. 2023: Create a New Situation of High-quality Education Development. [J]. Shanxi Education.2023.01.
2. Guo Meina. Problems and Countermeasures in the Management of University Books and Archives in the Information Age[J]. Shanxi Youth.2022.04.
3. Zhao Linlin. Management Strategy of University Books and Archives Based on Big Data Technology[J]. Qin Zhi.2022.09.
4. Yang Zhixue, Wang Jingjing. Construction of Big Data Cluster Based on Hadoop[J]. Information & Computer.2022.10.
5. Zhang Liping, Duan Shuping, et al. Design and Implementation of Big Data Processing Platform Based on Hadoop[J]. Electronic Test.2022.10.
6. Hong Piao. Research and Implementation of Search Engine Based on Lucene[D]. Dalian University of Technology.2016.09.
7. Lu Li. Research on the Application of MVC Design Pattern in JavaWeb Development[J]. Information & Communications.2020.04.
8. Liu Qinquan. Application of Improved TFIDF Algorithm in Text Analysis[D]. Nanchang University.2019.05.
9. Jia Xiaoxia. Design of Dynamic Retrieval System of Network Resource Index Information Based on Lucene[J]. Microcomputer Applications.2021.01.
10. Gao Jian. Design of Retrieval Method for Sci-tech Novelty Retrieval Based on Lucene Retrieval Tool[J]. Applications of IC.2022.04.