Journal of Contemporary Educational Research, 2021, Volume 5, Issue 12
http://ojs.bbwpublisher.com/index.php/JCER
ISSN Online: 2208-8474
ISSN Print: 2208-8466

Designing a Document Retrieval Method for University Digital Libraries Based on Hadoop Technology

Haixia He*

Beijing Normal University at Zhuhai, Zhuhai 519087, Guangdong Province, China

***Corresponding author:** Haixia He, 2398861328@qq.com

Abstract: With the development of big data, all walks of life in society have begun to venture into big data to serve their own enterprises and departments. Big data has been embraced by university digital libraries. The most cumbersome work for the management of university libraries is document retrieval. This article uses Hadoop algorithm to extract semantic keywords and then calculates semantic similarity based on the literature retrieval keyword calculation process. The fast-matching method is used to determine the weight of each keyword, so as to ensure an efficient and accurate document retrieval in digital libraries, thus completing the design of the document retrieval method for university digital libraries based on Hadoop technology.

Keywords: Hadoop technology; University digital library; Document retrieval method; Semantic similarity

Publication date: December 2021; ***Online publication:*** December 23, 2021

1. Introduction

In the context of the rise of big data, the traditional data storage and analysis system of university libraries has not been able to gain a competitive advantage. More and more university library readers turn to various network information resources. In the past few decades, the authority of university libraries has been challenged in providing academic resources and scientific knowledge information. Many university libraries are trying to change the methods and effectiveness of data collection. In addition, in the process of carrying out reader retrieval learning or training, university libraries have also established various databases one after another. The development of a data storage and analysis system for college libraries that can better meet the actual needs of readers and bring higher value to the services offered by university libraries [1]. However, at present, the large amount of available data and various resources in the network have increased the complexity of data collection; moreover, the cumulative growth of various relational and non-relational source data stored in local and cloud databases has caused universities to ponder on how to provide effective services for readers. This research integrates Hadoop technology based on big data into the document retrieval function of libraries to effectively solve the problem of semantic retrieval accuracy. The traditional document retrieval method uses a hierarchical retrieval model to judge information, which requires users to input retrieval fields with higher accuracy; otherwise, there would be a large amount of irrelevant data in the returned results [2]. In order to solve this problem, this study designs an efficient and accurate document retrieval method for university digital libraries based on Hadoop technology.

2. Introduction to Hadoop technology

Hadoop technology is a distributed computing environment based on the idea of map programming. The

principle of operation is as follows: decompose the task into multiple subtasks. These subtasks are assigned to different computing servers. Hadoop ensures the reliability of the results of each operation. When Hadoop maintains multiple copies of work data at the same time, it redistributes computing tasks to ensure the accuracy of literature retrieval [3].

Big data originated from Google. Google has proposed a set of infrastructure technologies based on distributed parallel clusters, using software capabilities to deal with frequent node failures in the cluster. The big data platform, Google, mainly includes four independent and tightly integrated systems: Borg Distributed Resource Management System [4]; Google File System (GFS) [5]; MapReduce programming model based on Google application characteristics [6]; distributed locking and large-scale distributed database, BigTable [6]. These big data processing technologies led by Doug Cutting were developed to support applications running in large clusters built on cheap commercial hardware, which is a key driving force for the rapid development of big data technology and applications [7].

Early versions of Hadoop, including Hadoop v1 and earlier versions, mainly consist of two core components: Hadoop Distributed File System (HDFS) and MapReduce. Among them, HDFS is an open-source version of GFS, while the computing framework implements the programming model of MapReduce, proposed by Google engineers [3]. There are also some open-source projects around Hadoop, which provide necessary support and supplements to improve the entire life cycle of big data processing. The commonly used software includes ZooKeeper, Hive, Pig, HBase, Kafka, Flume, Oozie, Mahout, etc. [3]. In May 2012, the alpha version of Hadoop v2 was released, and the most significant change was the addition of Yet Another Resource Negotiator (YARN) [3]. The emergence of YARN is to completely separate the computing framework and resource management, as well as solve the problems of poor scalability, single point of failure, and the inability to support multiple computing frameworks in Hadoop v1.

3. Analysis of the current situation of document retrieval in university libraries

3.1. Low efficiency of document retrieval in university libraries

Most university libraries have large collections of books. Therefore, it is difficult to find the necessary materials in a short period of time. The retrieval efficiency is not ideal, and it is difficult to solve the problem. One of the reasons is the unreasonable distribution of library documents, in which the arrangement order and subject classification are not clear enough, resulting in ambiguous search categories [8]; in addition, the books are not placed at their original positions upon return, and there are also some problems in management, which would be inconvenient for future searches, thus wasting a lot of time. Sometimes the search results would show that there are books; however, these books cannot be found; this reduces the search efficiency, and gradually students would be reluctant to use the search engine.

3.2. Flawed literature information retrieval system

Although the retrieval system brings convenience, if the system is flawed, the retrieval results would be greatly reduced. In addition, the flawed system would lead to untimely management of the warehouse, and contradictions may occur between the data found in the database and the library storage in the retrieval process [8].

3.3. Low quality of library management personnel

There are more or less problems with the low quality of library managers in various universities. If the comprehensive quality of managers can be effectively improved, the overall library management can be greatly improved. At present, China is a networked social environment, which requires networked office and retrieval. Therefore, managers need to improve their comprehensive level based on the actual situation

to avoid damage or confusion. In order to improve the efficiency of literature information retrieval, it is essential to attach great importance to the retrieval function of literature information, effectively improve students' personal awareness, fully integrate literature information retrieval into existing resources, and effectively improve the efficiency of resource utilization [8].

4. Application of Hadoop technology in university libraries for document retrieval

4.1. Hadoop technology in the document retrieval platform design of university libraries

In order to ensure that the Hadoop-based university library service platform designed in this study can achieve stable operation, in combination with the Hadoop distributed structure, the retrieval system structure of the university library service platform has been constructed. On this basis, it provides conditions for the subsequent operation of the platform. According to the retrieval needs of university library platform users, Hadoop with high reliability and good scalability in open-source organizations is used to store distributed library resource data. The platform is divided into four different levels, namely platform user access layer, various library service application layers, library basic resource data management layer, and library resource data storage layer [9].

At the platform user access layer, the public application port is mainly used for platform users to log in to their accounts. It provides conditions for subsequent users to enjoy various book resource information services on the platform. At the service application layer, components can be introduced into the platform by introducing cloud retrieval technology to provide different applications for different user services [9]. The basic resources management layer mainly manages the various Hadoop distributed files generated during the entire operation of the platform to ensure that various files can form a coordinated and unified working mode during the operation of the platform as well as achieve stronger resource retrieval functions for the system. The storage layer is a unified management of various data generated by the platform during the operation phase. The storage requirements for massive data in university library services can be achieved through network-attached storage (NAS) devices. NAS equipment is uniformly managed by the platform, using Hadoop distributed structure to realize the logical virtualization management of storage content and monitor the failure of various equipment during the operation of the platform to ensure that the platform is maintained before problems occur as well as ensure the security of platform user information and the stability of the platform [9].

4.2. Data mean clustering and library information resource retrieval output

In order to further improve the efficiency of literature retrieval in university libraries, it is necessary to consider the fuzzy logic-based clustering algorithm to divide the characteristics of the extracted library resource data and integrate the acquired data resources. Combined with the feature analysis method, the decision tree analysis of the nodes stored in the platform database is carried out, and MapReduce Architecture is introduced to complete full responsibility scheduling [10]. The task node defined by MapReduce serves as the execution node of the platform. It executes the instructions for the platform user during the operation of the platform. The task can be divided into a number of different subtasks, combined with different nodes of the task, to realize the retrieval instruction mapping and protocol verification of platform user retrieval requirements. Then, the platform storage function provided by HDFS is used to share all data nodes and provide required resource data for various service tasks during operation. At the same time, the parser can analyze and query various retrieval content proposed by the users, find the corresponding semantics in different query structures and query units, as well as use the metadata storage nodes in the platform to check the query structure. The metadata found in the query unit generates an execution plan. At the same time, the metadata storage node contains data information of various partition

structures of university libraries and can read and write various data in HDFS. The classification of library resource data is realized through the above cluster analysis. In combination with the output classification results of Hadoop, it can effectively provide the required search resources for the retrieval platform users in university libraries [10].

4.3. Semantic similarity calculation in document retrieval

After extracting the semantic keywords of the literature search, it is considered that the semantics contained in the literature and the semantics of the keywords are ambiguous. In this method, the Hadoop algorithm is used to calculate the semantic similarity in document retrieval. In the calculation process, keywords are used as text components. When calculating the similarity of text keywords of two matrix units, the calculation results of keywords under different topics are different. This is not a symmetric matrix. When the text contains multiple keywords, different keywords may be under different topics. The calculation process is also more complicated, which requires continuous improvement and perfection in subsequent data mining process [11].

In the document retrieval design, the document retrieval keywords must be extracted first, and then the semantic similarity can be calculated. In order to better meet the needs of students, a fast-matching method can be used to obtain the weight ratio of each topic keyword, taking into account the generation of different topics. As the document weight is different, it is necessary to build a document retrieval model. When the text is in the processing stage, the Hadoop algorithm of the text data source needs to be processed in advance to calculate the weight of the documents under the keywords one by one. The reference relationship between documents can be used to build an application network [12]. Each vertex in the network represents a document, and each edge represents a reference context. Hadoop technology can be used to change the prior probability distributions under different topics to obtain a directed network graph. Second, when the user generates a query, the user's search term can be a piece of contextual information. When the text is in the matching search stage, the system would preprocess the input long text, and based on Hadoop technology, it would perform matching queries on the subject keywords involved [13]. Considering that the relative weights of the documents under different keywords are different, the prior probability of each document needs to be calculated according to the needs of users.

5. Strengthen the utilization of Hadoop technology in retrieving literatures at university libraries

Reasonably optimize the library collection structure. Both paper books and e-books need to be properly optimized to keep the quantity and quality of resources at an ideal level; in addition, it is necessary to attract more readers to the library to read, regularly update the database, and improve the knowledge structure of readers at any time.

The importance of literature information retrieval should be strengthened. The retrieval of literature information is a very important work of the library. The library is one of the fundamental guarantees for providing services to readers. Regular updates can ensure the timeliness of teaching and research content. Therefore, the library should strengthen the importance of literature information retrieval. It should not simply consider the construction of collections and neglect the construction of periodical resources [8].

The accuracy of inquiries should be improved, retrieval automation should be appropriately strengthened, and convenience should be provided for the readers. It is also necessary to improve the overall quality of managers as the library staffs act as a "bridge" to ensure good communication between students and books or materials [8]. Therefore, the quality of the services they provide is very important. On the basis of improving the level of business, the sense of responsibility should be strengthened, and the employees should be given an opportunity to master a certain degree of computer operation. It is important to ensure

that they can handle part of the daily operations through the computer in a smooth manner. Finally, it is essential to ensure that the work of the library keeps pace with the times and maintains a harmonious environment.

6. Conclusion

In the context of the extensive application of "Internet +," the intelligent services of university libraries can effectively improve their intelligence, personalization, and initiative through big data mining technology, which greatly promotes the development trend of knowledge innovation; through big data mining, it is possible to obtain the knowledge contained in the library's big data, so as to meet the needs of intelligent services, such as knowledge sharing, recommendation, and navigation. The Hadoop technology-based literature retrieval method of university libraries provides convenient and accurate personalized services for many readers, improves readers' access and browsing speed, saves a lot of time in retrieval, and improves retrieval capabilities.

Disclosure statement

The author declares that there is no conflict of interest.

References

[1] Gong J, Yang H, Wen H, 2021, The Development and Realization of the Video Service Platform of Small and Medium-Sized University Libraries: Taking Nanchang Hangkong University "Live Library Network Service Platform" as an Example. China Education Information, 2021(07): 86-91.

[2] Shan Z, Shao B, 2021, Research on the Design and Evaluation of University Library Business Process Reorganization Based on the New Generation Service Platform. Research in Library Science, 2021(06): 27-35.

[3] Chen L, Huang J, Wang R, 2018, Overview of Hadoop Big Data Platform Security Issues and Solutions. Computer System Applications, 27(01): 1-9.

[4] Verma A, Pedrosa L, Korupolu M, et al., 2015, Proceedings of the Tenth European Conference on Computer Systems, April 21-24, 2015: Large-Scale Cluster Management at Google with Borg. EuroSys, Bordeaux, France, 18.

[5] Ghemawat S, Gobioff H, Leung ST, 2003, The Google File System. ACM SIGOPS Operating Systems Review, 37(5): 29-43. DOI: 10.1145/1165389

[6] Dean J, Ghemawat S, 2010, MapReduce: A Flexible Data Processing Tool. Communications of the ACM, 53(1): 72 -77. DOI: 10.1145/1629175

[7] Burrows M, 2006, Proceedings of the 7th Symposium on Operating Systems Design and Implementation, November, 2006: The Chubby Lock Service for Loosely-Coupled Distributed Systems. OSDI, Berkeley, CA, USA, 335-350.

[8] Shi S, 2017, Information Retrieval and Network Resource Utilization of Books and Documents in College Libraries. Exploration of Higher Education, 2017(S1): 139-140.

[9] Zhang Q, 2021, Design of University Library Service Platform Based on Hadoop. Science and Technology Innovation, 2021(23): 83-84.

[10] Zhang X, Zhang Y, Yang F, 2020, Research and Design of Document Retrieval Method for University Digital Library Based on Hadoop Technology. Microcomputer Applications, 36(07): 11-13, 23.

[11] Liu F, 2017, Research on Reading Promotion of University Libraries Based on 4I Marketing Principles. Library Work and Research, 1(9): 36-39.

[12] Han Z, Luo R, 2017, Experimental Research on the Influence of Academic Users' Emotional Control and Mental Models on Information Retrieval Performance. Information Theory and Practice, 40(1): 59-64.

[13] Zhou D, Zhao W, 2017, Research on Result Reordering in Personalized Cross-Language Information Retrieval. Computer Engineering and Science, 39(10): 1922-1929.