Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6207054, 11 pages
https://doi.org/10.1155/2022/6207054

Research Article

Analysis of the Path of Utilizing Big Data to Innovate Archive Management Mode to Enhance Service Capability

Wenjingling Luo and Haijian Hu

Guangdong Polytechnic of Science and Technology, Zhuhai, Guangdong Province 519090, China

Correspondence should be addressed to Wenjingling Luo; 201522600090@mail.bnu.edu.cn

Received 11 July 2022; Revised 9 August 2022; Accepted 23 August 2022; Published 11 October 2022

Academic Editor: Kuruva Lakshmanna

With the rapid development of modern science and technology, we are now in an era of big data and digitalization of network, and people's normal work and life are also implicitly influenced. Archives, as the information of various work examinations, are the imprints of the past and the basis for guiding the future. Compared with the traditional management, the information-based archive management is more efficient and of better quality. The efficiency of all sectors of society in the background of the information age has taken a qualitative leap and improved. In order to develop the archive management business and make the archive management more efficient and secure, a set of scientific and efficient information-based archive management work mode should be worked out on the road of practice in the new era. With the arrival of the era of big data brings fundamental changes to data mining technology, which can analyze and process a large amount of data and dig intelligently, making it possible to dig deeper into the data. Therefore, it is necessary to actively explore archival data mining technology, so that data mining technology can be widely applied to archival management work, and the potential archival information can be excavated and delivered as much as possible to play its proper role and maximize the economic and social benefits of archival data. The article analyzes the characteristics of archive management in the context of big data, proposes a specific path to improve the quality of archive management work in view of the current problems in archive management work, and designs the archive management structure based on big data for improving the quality of archive management work.

1. Introduction

Archives play an important reference role in the process of social and economic development. With the advent of the era of big data, big data technology has been widely used in all aspects of society, which not only improves the efficiency of archive management but also contributes in the innovation of work mode of archive management. It further improves the quality of archive storage and efficiency of archive utilization. From the current state of archive management work, although big data technology has been initially applied in archive management, but in practice there still exists the problem of promotion and application is not in place, there is still the phenomenon of using the traditional archive management work model, resulting in low efficiency of archive management; it is difficult to adapt to the requirements of archive management work in the era of big data [1–3]. The concept of archive management is backward. Under the current widespread promotion of big data technology, the data structure of archival information resources has undergone fundamental changes, and the quantity and types of archival information data have increased, forming a huge database of archival information, and the deep value mining of these archival information resources requires advanced archival management concepts. In archive management practice, archive managers carry out the development and maintenance of archive cloud storage system on the cloud platform to develop archive information data with utilization value, which can provide remote archive information services for users and effectively improve archive work [4].

However, due to the current archive management practice, some archive management departments and archive managers are still following the traditional archive management model, and the archive management concept is rather backward, especially the big data technology application skills are backward, which makes it difficult for archive management to effectively meet the realistic needs of archive users and is not conducive to the realization of the modernization goal of archive management. The archive management facilities are insufficient, and archive users must have perfect archive infrastructure to get high level archive services. Under the background of big data, archive managers must rely on basic service platforms such as cloud platforms to better complete the storage and utilization of archival information data in the process of archival information services. However, some archive management departments fail to equip corresponding archive management infrastructure, resulting in the lack of infrastructure support for archive management, which not only prevents them from completing the collection and arrangement of archive information data but also prevents them from categorizing and storing in a timely manner, not to mention providing remote archive information services to users. In short, the lack of perfect infrastructure service facilities affects the comprehensive integration of archival information data with big data in archival management departments and blocks the improvement of archival management work in archival management departments. Poor security awareness in management work and the security of archival information resources are the primary goals of the archival management system. Under the background of big data, archive management work needs to be carried out by applying various advanced big data technologies. In this process, it is necessary to take archival information data security as the primary consideration, improve the awareness of data security management, ensure the security of archival information database, and prevent archival information data from being leaked or tampered with. However, some archive management departments are backward in archival information data security management technology due to the lack of corresponding data security prevention and control technology, and some archive managers lack corresponding awareness of archival information data security management, which makes it difficult to provide secure archival information services to archive users [5–7].

It greatly increases the dynamism of archival information data. In the era of big data, data storage replaces the traditional paper archive management mode, which to a certain extent expands the content and space of the original archive management. Archival information data storage does not occupy substantial storage space, thus making the collection and arrangement of archival information data no longer restricted by time and space, theoretically making archival information data storage reach infinity, and also enabling the dynamic collection, arrangement, and storage of archival information [6–8]. For example, for personnel record management, there is a large amount of dynamic archival information data to be collected, organized, and stored in the process of personnel recruitment, interview, entry and exit, etc. Under the original record management mode, only a limited number of key archival information can be recorded due to space limitation. However, under the background of big data, these archive management data can be collected and recorded as much as possible, which improves the dynamics of archive information data and thus makes archives management more comprehensive and real.

Obviously increasing the complexity of archival management, in the era of big data, due to the rapid increase of the total amount of archival information data, the types of archival information data are also more diversified, although it can better meet the different needs of people in different industries, but to a certain extent, it also increases the complexity of archival management, especially when classifying archival information data; if the accuracy of archival information types cannot be ensured, it may lead to confusion and increase the complexity of archival information data. If there is a lack of effective archival information analysis tools, it may lead to some data with value and some data without value in the face of the ponderous archival information data, which will generate a large amount of useless information data when archival information data is output, and undoubtedly increase the difficulty of archival management [9–11]. The application of big data technology in archive management has a great impact on the traditional way of collecting and organizing archival information, and the source channels of archival information data are more diversified, which enriches the types of archival information data, some of which are structured data and some are unstructured data. The biggest difference between it and traditional paper archives is that big data technology can not only keep the original paper archives but also store the archival information data such as audio and video, thus breaking through the traditional paper archive management mode. At the same time, after storing these different types of archival information data, it can also transform the types of different types of archival information data according to the needs of archival management and dig deeper into the already classified archival information to find information that is more beneficial to their needs, thus improving the convenience of archival management.

2. Related Work

2.1. Innovative Archive Management Mode. Archive management is a part of the important work of enterprises. Innovative archive management methods and optimized archive management workflow in enterprises can promote the orderly development of enterprises. Therefore, enterprises should actively innovate the way of enterprise file management and combine the needs of social development, apply information technology in file management to meet the needs of social development, create an information-based office atmosphere, and promote the innovative development of file management [12–15].

The necessity of actively carrying out the innovative mode of archive management service in the information era has raised the importance of history and humanistic literacy in the continuous development of society, and archive management contains a variety of traditional historical materials, among which are history, humanistic literacy, folk customs, etc. Archive management is an important way to store data materials in the development of the times, which is of great benefit to the inheritance and development of Chinese culture. At the same time, there is a wealth of library information hidden in the library archive management, which are valuable original materials and have a very precious collection value [16–18]. The library covers rich knowledge and culture and is the main place for cultivating talents and inheriting culture and science, so the library should pay attention to innovation, build an information-based office model in the library, innovate the search function, scientifically obtain thematic resources, and optimize the archive management service system, which is beneficial to the innovative development of society.

In the continuous development of society, the content of library archive management work is gradually increasing, and science and technology are constantly being developed and applied, and in this environment, the matters and contents of archive management work in libraries are expanding, and their service forms are also expanding. Therefore, it is necessary to innovate the service mode of archive management in libraries, which can not only give full play to the value of library archive management but also play the role of guidance and education. The use of technology to reform archive management services can effectively spread Chinese culture. The lack of corresponding archival management infrastructure has led to the lack of infrastructure support for archival management, which makes it impossible not only to complete the collection and arrangement of archival information data but also to categorize and store them in a timely manner and to provide remote archival information services for users. In short, the lack of perfect infrastructure service facilities affects the comprehensive integration of archival information data with big data in archival management departments and blocks the improvement of archival management work in archival management departments.

It is necessary to take archival information data security as the primary consideration, thereby improve the awareness of data security management, ensure the security of archival information database, and prevent archival information data from being leaked or tampered with. However, some archive management departments are backward in archival information data security management technology due to the lack of corresponding data security prevention and control technology, and some archive managers lack corresponding archival information data security management awareness, which makes it difficult to provide secure archival information services to archive users.

It is necessary to establish the awareness of big data management. To closely combine big data technology with archive management, we must first establish the awareness of big data management. It is necessary to change the mindset of archive management from the ideological point of view. The service consciousness has a direct impact on the archive management service [19–21]. Apply big data technology in archive management work, combine the actual needs of archives users, strengthen the big data thinking and archive service consciousness of archive management personnel, build and establish modern archive management mode, especially to consciously transform their own archive service consciousness, change from traditional archive management consciousness to big data archive management consciousness, strengthen the application of big data technology in daily work, and consciously improve archive service. The ability of archival services should be consciously improved. It is necessary to strengthen the construction of remote service capacity for archival information and data. The archive management department should strengthen the management of archival information data, improve the accuracy of archival information retrieval through the application of modern information technology in the archive management system, such as network technology, information retrieval intelligent technology, GIS technology, and other advanced technologies, and enhance the information management efficiency and service quality of the archive management department through the construction of the archive network remote service platform.

2.2. Big Data Technology. In recent years, with the development of cloud computing and mobile Internet, the global data volume has been increasing. The most important characteristic of big data is that the data has high dimensionality and structural complexity, and usually, these data are accompanied by a large amount of interference and noise data. As a result, algorithmic models that once performed well for small and low-dimensional datasets will no longer be directly applicable to the situation of today's big data. To address this problem, many scholars and researchers have started to work on parallel data mining and archival classification algorithms. Apache Spark has implemented a series of machine learning algorithms in an in-memory-based computing model, mainly through the HDFS distributed file system to increase the throughput and concurrent data access. Experiments have proven that Spark can compute 10 times faster than Hadoop reading and writing data blocks from memory and 100 times faster than Hadoop reading and writing from disk [21–24]. Also, Spark provides more than 80 easy-to-use operation operators, making it easier to write parallel programs. The entire ecosystem of the current Spark big data computation engine and its main functional modules are described in Figure 1.

As the big data technology system continues to mature, the internal technology composition continues to differentiate, from the core technology facing the needs of storage, processing, and analysis of massive data, to the supporting technologies of data management, circulation, and security, gradually forming a clear hierarchy and complete division of labor of the big data technology system. The data foundation technology responds to a variety of data characteristics. For big data volume, data sources heterogeneous and diverse, data timeliness, and other characteristics have given rise to the efficient completion of a large amount of heterogeneous data storage and computing technology needs. Under this demand, the traditional centralized computing architecture has emerged as an insurmountable bottleneck, and the storage and computing performance of traditional relational database is limited, so distributed storage and distributed computing frameworks have emerged [17, 25–27]. The study in [28] implemented distributed computing for the recovery of original individual

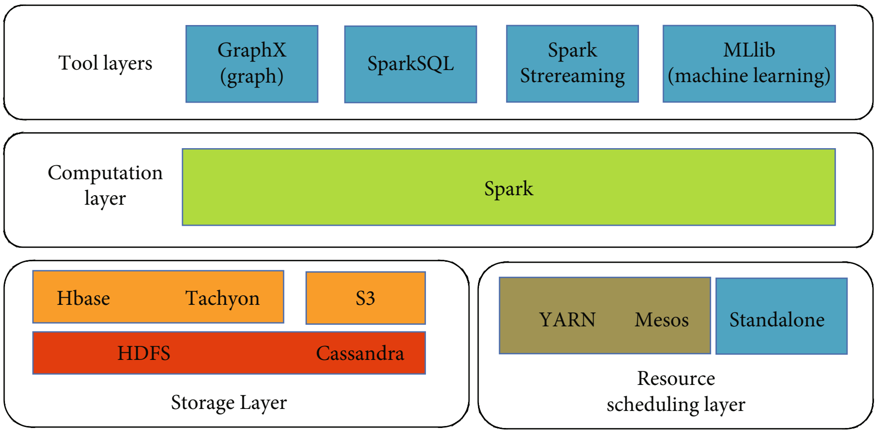

Figure 1: Ecosystem and functional modules of big data processing framework.

personal data (IPD) from empirical IPD summaries. The main objective of the study is to reproduce original features of IPD in the artificial data in order to recover the IPD inferences. The random-effect logistic regression and Gaussian copula were used for the said purpose. The study in [29] used the concept of "coded computing" that used coding theory to inject data and reduce issues pertinent to data redundancy and removal of bottlenecks in large scale distributed computing. As part of the MapReduce-based distributed computing structure, a coded distributed computing scheme was implemented that reduced communication load to shuffle the intermediate computation results.

For massive structured and unstructured data batch processing, distributed batch computing frameworks based on Hadoop, Hive, and Spark ecosystem have emerged; for the demand of real-time computing feedback of time-sensitive data, distributed stream processing computing frameworks such as Storm, Flink, and Spark Streaming have emerged. Data management technology improves data quality and availability. After the relatively basic and urgent data storage and computing needs have been satisfied to a certain extent, how to manage and precipitate data has become a major demand. Due to the long chain and high complexity of data generation within enterprises, but the general lack of effective management, there are often problems such as difficult data access, low accuracy, poor real time, and confusing standards, leading to numerous obstacles to the subsequent use of data. In this case, data integration technology for data integration and data management technology for realizing a series of data asset management functions have emerged.

Data analysis application technology is extremely predominant and beneficial in tapping data value. To carry out data analysis and excavate data value, including statistical analysis and visualization presentation technology represented by BI tools, as well as traditional machine learning and deep learning based on deep neural networks, excavation analysis and modeling technology have emerged to support the excavation of data value and further apply the analysis results and models to actual business scenarios. Data security circulation technology helps secure and compliant data use and sharing. As the value of data is explored, data security issues are becoming more and more prominent, with data leakage, data loss, data misuse, and other security incidents popping up all the time. Data protection technologies such as access control, identity recognition, data encryption, data desensitization, and privacy computing are actively evolving in the direction of more adaptive to big data scenarios [30, 31].

3. Methods

3.1. Model Architecture. With the development of the Internet, the era of big data is gradually coming, and the Internet is full of all kinds of information; most of these resources are in the form of this paper; how to effectively handle and organize this information becomes especially important. A large amount of data can improve the accuracy of models, and complex machine learning algorithms take a lot of time, so there is an urgent need for such key technologies as distributed and parallel computing. The combination of big data and machine learning opens a new path for archival classification, so archival classification systems based on big data become very meaningful.

The archival system based on big data is mainly divided into model application layer, data mining layer, data preprocessing layer, data storage layer, and data collection layer. The model application layer calls the interfaces provided by the data processing layer to realize archive classification and provide services for users. In the whole system, the lower layer application provides services for the upper layer application. Its main processing flow is that the data acquisition layer collects data, stores data using the services provided by the data storage layer, and preprocesses the stored data by the data preprocessing layer. The data processing layer mainly uses the powerful computing power of Spark and the algorithm to mine and calculate the data conforming to the algorithm data format, get the data classification model, and finally apply the data model to provide services for users. The system architecture is shown in Figure 2.

3.2. MapReduce. In the MapReduce framework, the Shuffle module is the bridge between the Map phase and the Reduce phase. The Reduce phase is responsible for pulling the corresponding fragment data from each map node to the reduce

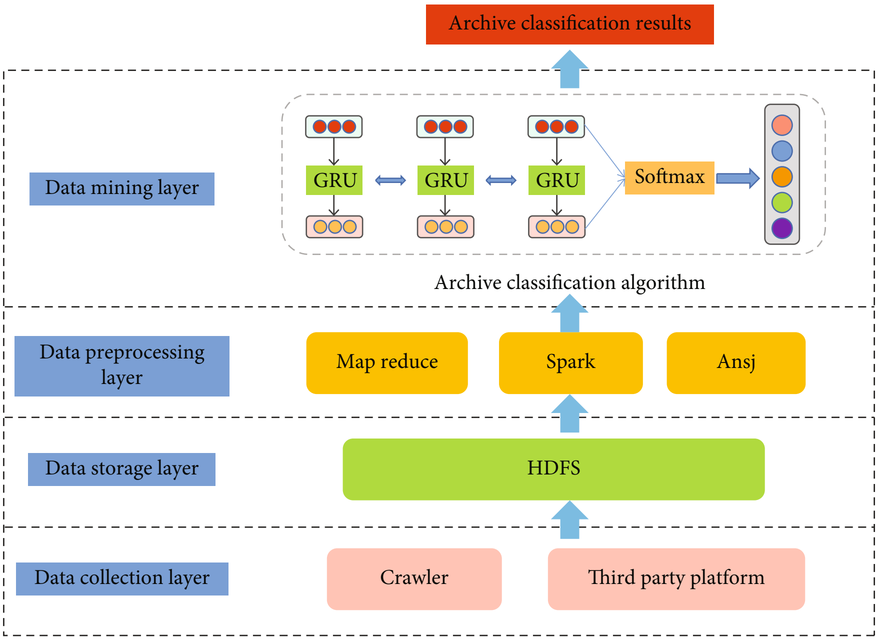

Figure 2: Model structure.

node, sorting, merging, and finally calculating the result. In the whole shuffle process, because the reducer side needs to read the intermediate results output from the map stage from different nodes, so it often involves a lot of disk reading and writing and network data transmission, so the performance of the shuffle also directly determines the performance of the whole program. In other words, if the performance of shuffle is improved a little bit, the computational performance of the whole system will be greatly improved. The major advantage of MapReduce is that it helps in processing data in parallel processing and uses commodity clusters for the distribution. MapReduce is highly scalable and reduces the cost of storage and processing of data in order to meet the growing data requirements. There are various successful applications of MapReduce framework. As an example, the study in [32, 33] used a Hadoop-based MapReduce framework for implementing an improved chaotic image encryption algorithm for massively remote sensed images in case of parallel IoT applications. The study explored the parallel methods of image encryption for large number of remotely sensed images in Hadoop, and the results highlighted the efficiency and scalability of the approach.

Spark, as a new generation of big data processing engine, is also an implementation of MapReduce framework, so it will also have its own shuffle implementation process. The Shuffle phase in spark is divided into two processes: *ShuffleWrite* and *ShuffleRead*, in which the *ShuffleWrite* process mainly sorts the intermediate results output by *MapperTask* by *paritiionId* partition number and key primary key and then writes all the ordered records to the disk of the node where the current task is located by Flush overflow. The *ShuffleRead* phase is when all the *MapperTasks* have finished executing, and each ReducerTask, according to its task ID number, goes to a different node to pull the data that belongs to its own output in the previous *ShuffleWrite* phase. Then, they sort and aggregate the data, then compute it, and finally return the result to the client. However, in the *ShuffleWrite* stage, when executing *MapTask*, because of the large amount of input data, the data can only be persisted to disk while calculating, i.e., only one source input data is read at a time and then calculated, and then, the calculation results are stored in the memory data structure *PartitionedAppendOnlyMap*, which is essentially a data array, occupying a contiguous section of memory space, after calculating one piece of data, then continuing to read the next piece of data, and then putting the calculation results into memory.

The system process is shown in Figure 3. The collected big data is stored in the distributed file system of HDFS system and preprocessed by MapReduce batch data; then, Spark's powerful computing power trains the data according to the file classification algorithm, generates the file classification model, and finally encapsulates the file classification model into a service and provides it to the users.

3.3. File Archive Data Processing. To realize Chinese multilabel archive classification, Chinese multilabel archive big data is established, and the semantic similarity features of Chinese multilabel archive big data are extracted by combining the method of semantic information fusion with the method of fuzzy partition block fusion, and the partition block is stored and classified by combining the copy detection method. The statistical fusion of Chinese multilabel archive big data is carried out by the method of context mapping. First, the ontology structure mapping is performed on Chinese multilabel archive big data, and the mapping model is defined as

$$g(x) = \frac{1}{1 + e^{-x}}, \tag{1}$$

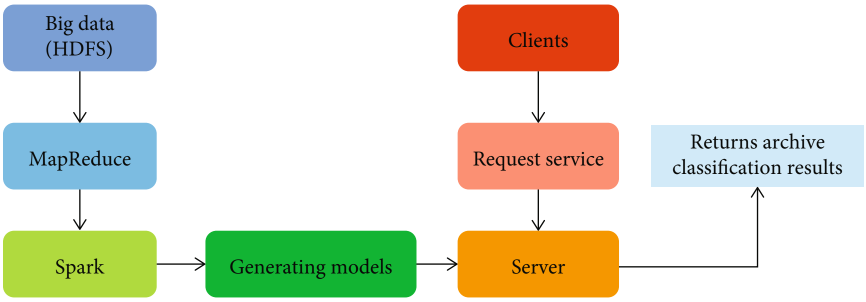

Figure 3: System flow chart.

where e^{-x} denotes the similarity of semantic ontology fusion on Chinese multilabel archive big data, and the expression is calculated as

$$e^{-x} = \frac{1}{1 + e^{-\Theta^{\tau} z}}. \tag{2}$$

According to the above formula, the fuzziness coefficients of Chinese multilabel archive big data are extracted, and the semantic mapping relationship is established to obtain the association rule set of Chinese archive big data, and the output is

$$s(t) = \text{sigmoid}[Uw(t) + W_x g_S(t-1)], \tag{3}$$

where U denotes the knowledge of the relationship between elements of Chinese multilabel archive big data; g denotes the feature weight of Chinese multilabel archive big data; $w(t)$ is the feature distribution coefficient of semantic information; and W_x denotes the fuzziness function of Chinese multilabel archive big data, and establish the semantic mapping relationship, and adopt a quadratic to represent the Chinese multilabel archive big data to establish a semantic mapping relationship distribution model, where E_i, E_j are the entity sets of Chinese multilabel archive big data, d is the similarity information of Chinese multilabel archive big data, and t is the sampling interval. According to the above analysis, combining the semantic ontology feature reconstruction method with the archival information classification fusion processing method, the Chinese multilabel archival big data analysis model is established as

$$Q = s(t) * g(x) + \sum_{i=1} \left\{ \left(E_i, E_j, d, t \right) \right\}. \tag{4}$$

The Chinese multilabel archive big data is analyzed according to the Chinese multilabel archive big data analysis model, and the feature extraction of Chinese multilabel archive big data is carried out based on the analysis combined with the knowledge of the relationship between elements.

3.4. Feature Extraction. Based on the obtained Chinese multilabel archive big data analysis results, the statistical feature detection of Chinese multilabel archive big data is carried out by combining the knowledge of the relationship between elements and logical inference methods, and the Chinese multilabel archive statistics containing semantic attribute feature quantity and sample autocorrelation feature quantity is obtained, which is expressed as

$$E = \frac{1}{QN} \sum_{N} \sum_{c=1} \left(f(x, w) - y_c \right)^2, \Delta w = \frac{\partial E}{\partial w}, \tag{5}$$

where y_c denotes the coefficient of Chinese multilabel archive feature distribution. On this basis, the statistical distribution set of Chinese multilabeled archive big data is constructed, and the semantic attribute feature quantity of Chinese multilabeled archive big data is obtained using description logic as

$$E^* = \frac{pa_n}{p(a_n - 1)} * \sum_{n=1} (pd_n - 1 + b). \tag{6}$$

In domain knowledge and structural knowledge, Chinese multilabel archive feature fusion is performed, and information fusion is performed on two heterogeneous ontologies to obtain the overlapping distribution set of Chinese multilabel archive big data defined as $D = \{S.(t), T.(t), U(t)\}$, where $S.(t)$ denotes the relevant domain knowledge set of Chinese multilabel archive big data and $T.(t)$ denotes the predicate logic distribution set, by pure archive big data scheduling; establish the segmented sample test model of concepts, instances, and attributes, and get the sample autocorrelation feature quantity as

$$\omega_j = \sum_{j=1,k=2} \omega_{k-1,j} * D^T. \tag{7}$$

The autocorrelation feature distribution set of Chinese multilabel archive big data is established, and the linear regression analysis method is used to obtain the triadic form of Chinese multilabel archive big data as

$$M = \sum_{I=1}^{C} \frac{\omega_j}{\mu_{ik}}, k = 1, 2, \cdots, n, \tag{8}$$

where k is the fuzziness coefficient of Chinese multilabel archival big data, and the discrete sequence scheduling method is used to construct the semantic feature extraction model of Chinese multilabel archival big data, which is

Table 1: System hardware environment configuration table.

Configuration name	Configuration type	Configuration specifications
CPU	Intel Xeon E5-2682v4	2 cores
Memory	Unknown	4 G
Disk	Efficient cloud drive	20 G

expressed as

$$F* = \frac{1}{n}\sum_{j=1} M\left(\omega_j + E^*\right)^T. \tag{9}$$

3.5. Archive Classification Algorithm. The extreme learning machine (ELM) is a forward neural network with only one single layer. The use of ELM has several advantages which includes simplicity in its implementation, enhanced speed of learning, and finally its applicability in nonlinear kernel functions. The model generates good generalization performance and learns extremely faster than traditional back propagation networks. ELM has been predominantly used in various applications. As an example, it has been implemented in deep learning framework for the prediction of financial market. A hybrid model is developed in [34] using auto encoder and kernel ELM for the prediction of financial market. Similarly, the study in [35, 36] proposed a heuristic Kalman filter-based optimized ELM framework that helped in the prediction of useful life of the capacitors. The ELM was preferred in the study as it worked with enhanced speed enabling efficient forecasting of capacitor life. The working principle of ELM is different from that of traditional neural networks, such as the weights of BP neural network are obtained by gradient descent algorithm, while the weights of ELM are obtained by parsing expressions without intermediate iterative computation, and ELM runs faster. Let x denote the value of the input sample, and the threshold, weight, and node of the hidden layer are a_i, b_i, and L, respectively; then, the output value of ELM is

$$f_L(x) = \sum_{i=1}^{L} G(x, a_i, b_i)\beta_i, \tag{10}$$

where G is the activation function of the hidden layer, and the radial basis function is used in this paper, as follows.

$$G(x, a_i, b_i) = g(b_i x - a_i) = \mathrm{e}^{-b_i x - a_i}. \tag{11}$$

ELM divides all archival data into m categories by the activation function of the hidden layer and uses $h(x)$ to represent G.

$$f_L(x) = \sum_{i=1}^{L} h(x)\beta_i. \tag{12}$$

4. Experiments and Results

4.1. Experimental Setup. The system hardware environment system is deployed based on Spark cluster, and the whole cluster consists of three servers. The configuration of each server of the cluster is consistent, and the configuration information is shown in Table 1.

Spark service is deployed on the system cluster, and Spark service also has many dependent services such as Java environment and Scala environment. The deep learning system implemented in this paper is based on the Python environment, so it should also follow the Python service. The software configuration information is shown in Table 2.

The experimental data comes from the Chinese news archive management agency, and the database contains many Chinese articles from news websites, many of which are news reports written by editors of the websites and contains 9833 Chinese documents. Because there are some topics in the dataset, the document collection contains a much higher number of documents than the document collections of other topics. The training process performance enhancement and loss convergence are shown in Figures 4 and 5.

4.2. Experimental Results and Analysis. In the experiments, 60% of the preprocessed dataset is taken as the training set and the remaining 40% as the test set by random sampling method. First, we obtain an optimal coefficient a from the Spark computing environment. In this section, we conduct experiments using different values of coefficient a on the four datasets mentioned above and determine the impact of different coefficients on the final classification accuracy by comparing the accuracy of the experimental results. In these comparison experiments, we set the values of the coefficients a between (0,1) and incremented from 0 to 1 on the axes, with each increment being 0.1. Figure 6 shows that the accuracy of the experimental results on the four datasets increases with each increment of the coefficient a. Clearly, the accuracy of the classification reaches a minimum when the coefficient $a = 0$. From the experimental results, the optimal value is obtained when the coefficient $a = 0.9$. Therefore, all the next experiments are performed on the basis of the optimal coefficients.

In addition, the experiments in this section compare the effect of Spark before the improvement of the memory prediction algorithm in the shuffle stage, and the experiments show that the improved Spark achieves a higher improvement in computational efficiency than before the improvement. The K-means algorithm is the most classic division-based clustering method and is one of the top ten classical data mining algorithms. K-means algorithm is based on the basic idea of clustering k points in space and categorizing the objects closest to them. By iterative method, the value of

Table 2: System software environment configuration information table.

Software type	Software version
Operating system	Ubuntu 14
Virtualization services	Docker
Spark services	Hadoop3.2.1 + Spark2.3.4
System dependency services	SSH + Scala2.11 + JDK1.8 + Python3.6 + Redis3.2.100

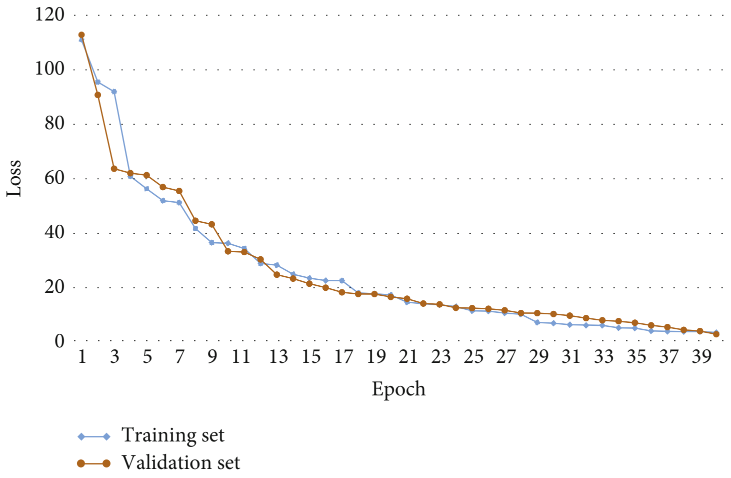

Figure 4: Schematic diagram of training process performance improvement.

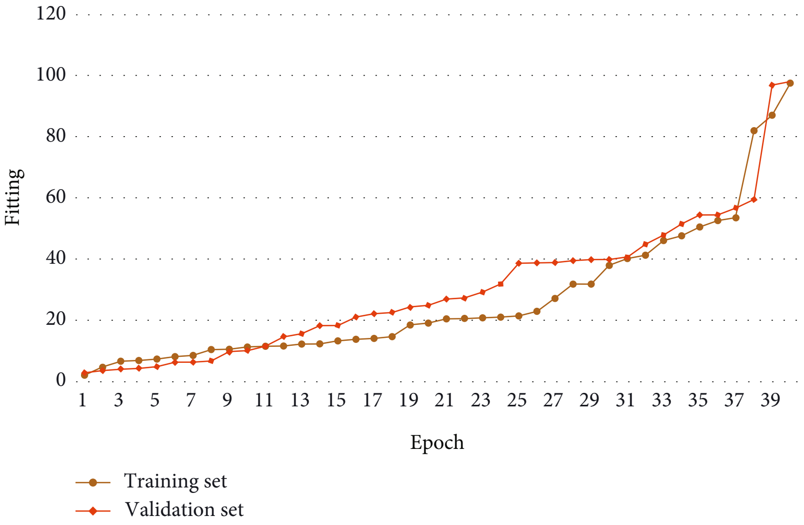

Figure 5: The training process loss convergence schematic.

each clustering center is updated one by one until the best clustering result is obtained. Bayes is a classification model in machine learning, which is widely used in practice due to the simplicity and efficiency of the algorithm. We will compare the job execution time of this benchmark on Spark and Spark+. The experimental results are shown in Figures 7 and 8.

The test results can be seen from the figure that the average time for classification reaches 4.98 min when running on the unoptimized spark, while the improved spark, i.e., with the introduction of the new shuffle memory prediction algorithm, reaches an average speed of 3.28 min, a speedup of 34.73%. After we finished the work of classifying archives with deep learning model in standalone mode above, the next step is to test the performance of the archive's classifier based on Spark neural network designed and completed in this paper. The model is created using the built-in distributed separator DisTextCNN, which is a distributed version

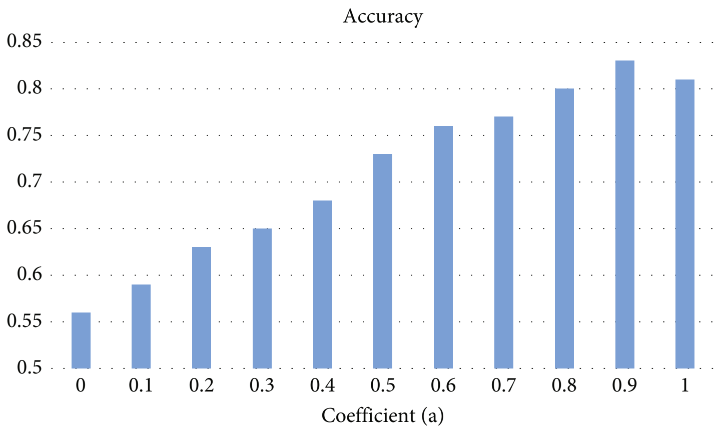

Figure 6: Trend graph of accuracy with different coefficients.

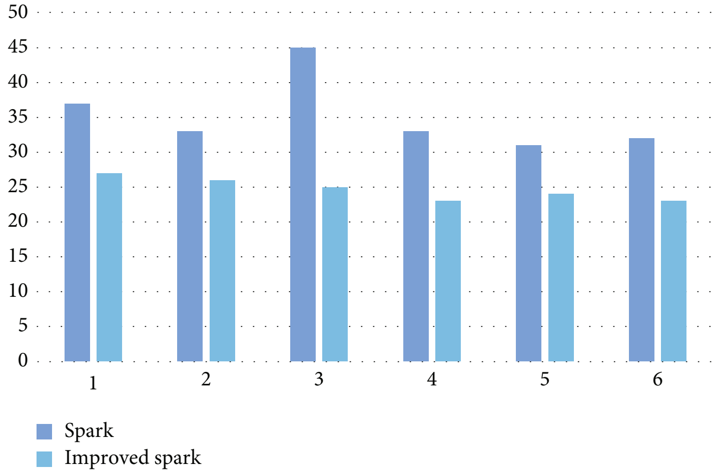

Figure 7: Comparison of execution time (seconds) of K-means on Spark before and after improvement.

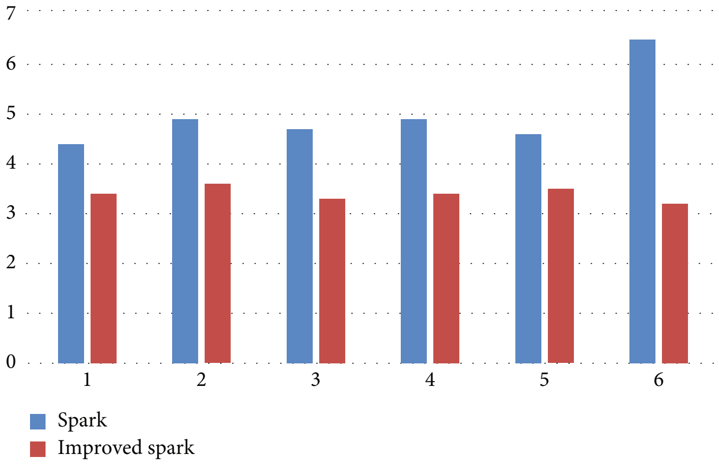

Figure 8: Comparison of execution time (minutes) of Bayes algorithm on Spark before and after improvement.

of the deep learning model TextCNN in stand-alone mode implemented in this system. The training and validation sets are divided into 70% and 30% and are stored in the system's HDFS. The parameters of the distributed model are shown in Table 3.

The paper is an extension of Spark with the intention of distributed training of deep learning models to improve efficiency, so the performance comparison test between distributed and stand-alone models is conducted. The training of TextCNN and DisTextCNN is completed above, and the training data of

Table 3: Experimental platform and configuration.

Models	Training set accuracy	Validation set accuracy	Training time/min
TextCNN	0.90	0.80	87
DisTextCNN	0.92	0.80	55

the two models are obtained. The number of iterations of the distributed model is twice that of the standalone model, because it includes external iterations, although the number of iterations is more than that of the standalone model; the training time of the distributed model is one-third less than that of the standalone model. Performance test results using the distributed deep learning archival classifier system show that the creation of models, model training, and classification tasks are very simple for users to achieve ease of use. By training the distributed archive classification model created in the system, the accuracy of the model reached the target value of 80%, and comparing the performance of the distributed model with that of the standalone model, it was found that the training speed of the distributed model was significantly higher than that of the standalone model, achieving the high efficiency required by the system.

5. Conclusion

In the context of the big data era, the competition for data resources is becoming more and more intense, and it is important to fully recognize the great driving effect of data on the development of human society. Big data technology should be fully utilized in archival work, and archival materials should be deeply mined so that they can play their proper functions in the big data environment and promote the progress and development of society. The emergence of digital archive management has improved the efficiency of archive management, solved the problem of unreasonable application of archival information, and promoted the development of enterprises and institutions to a certain extent. In the future construction of the value chain of archive management, the effect of digital archive management can be ensured through innovative management concepts, expansion of service contents, and establishment of management teams. The wide application of big data technology in the archive management work has put forward new target requirements for the archive management work in the new era. The archive management departments must address the characteristics of archive management in the context of big data, fully establish the awareness of big data management, innovate and improve the new archive management system, establish a wisdom archive management platform, carry out two-way data management of archive information, and comprehensively improve the level of archive management. Although the proposed ELM-based framework yields promising results, the traceability and interpretability of the results still remain a challenge. As part of future study, explainable AI techniques could be considered to achieve the same and achieve more confidence on the generated results.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Innovation Team Fund Project of Regular Colleges and Universities of the Education Department of Guangdong Province: “Research and Innovation Team of Citizenship and Morality Education in Guangdong-Hong Kong-Macao Greater Bay Area” (No. 2020WCXTD023) and the Teaching Reform Project of Guangdong Polytechnic of Science and Technology: “Research on the Path and Influencing Factors of Improving the Teaching Efficacy of Ideological and Political Course Teachers” (No. JG202232).

References

[1] S. Ding, S. Qu, Y. Xi, and S. Wan, “A long video caption generation algorithm for big video data retrieval,” *Future Generation Computer Systems*, vol. 93, pp. 583–595, 2019.

[2] Z. Gao, Y. Li, and S. Wan, “Exploring deep learning for view-based 3D model retrieval,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1, pp. 1–21, 2020.

[3] Z. Gao, H. Xue, and S. Wan, “Multiple discrimination and pairwise CNN for view-based 3D object retrieval,” *Neural Networks*, vol. 125, pp. 290–302, 2020.

[4] S. Ding, S. Qu, Y. Xi, and S. Wan, “Stimulus-driven and concept-driven analysis for image caption generation,” *Neurocomputing*, vol. 398, pp. 520–530, 2020.

[5] J. Pei, Z. Yu, J. Li, M. A. Jan, and K. Lakshmanna, “TKAGFL: a federated communication framework under data heterogeneity,” in *IEEE Transactions on Network Science and Engineering*, p. 1, 2022.

[6] M. Zhao, C. Chen, L. Liu, D. P. Lan, and S. Wan, “Orbital collaborative learning in 6G space-air-ground integrated networks,” *Neurocomputing*, vol. 497, pp. 94–109, 2022.

[7] S. Magomedov and A. Lebedev, “Protected network architecture for ensuring consistency of medical data through validation of user behavior and DICOM archive integrity,” *Applied Sciences*, vol. 11, no. 5, p. 2072, 2021.

[8] L. Lau, “Leadership and management in quality radiology,” *Biomedical Imaging and Intervention Journal*, vol. 3, no. 3, 2007.

[9] L. Jia, W. U. Jianhua, and L. U. Jiang, “Research on the method of archival business process re-engineering in the view of big data,” *Archives Science Study*, vol. 3, 2018.

[10] M. Fafchamps and M. Soderbom, “Wages and labor management in African manufacturing,” *Development and Comp Systems*, vol. 2, 2004.

[11] M. O. Cibarolu and B. Yalnkaya, “Belge ve ariv ynetimi sürelerinde büyük veri analitii ve yapay zeka uygulamalar,” *Information Management*, vol. 2, no. 1, pp. 44–58, 2019.

[12] P. Comitz, S. Ayhan, G. Gerberick, J. Pesce, and S. Bliesner, "Predictive analytics with aviation big data," *IEEE*, vol. 1, pp. 1–35, 2013.

[13] E. Alexopoulos, F. Charizani, D. Markea et al., "Estimation of the effectiveness of the waste management plant of a big industry," *Epitheorese Klinikes Farmakologias kai Farmakokinetikes*, vol. 23, no. 1, pp. 175–182, 2005.

[14] B. C. Cho and H. S. Yuk, "The role of archive as cultural memory in the age of big data," *Geographical*, vol. 12, no. 2, pp. 1–10, 2014.

[15] K. Hallam, "Wrestling with risk. Risk-management award winners find that a little effort can lead to big rewards," *Modern Healthcare*, vol. 28, no. 43, pp. 56–8, 60, 1998.

[16] B. Shankar, "Analytics is new mantra in data management," *Chinese Nursing Research*, vol. 40, no. 5, pp. 16-17, 2010.

[17] H. Chen, J. Xie, S. J. Wang et al., "Research on intelligent management system of meteorological archives based on big data framework," *Advances in Data Science and Adaptive Analysis*, vol. 13, no. 3n04, 2021.

[18] S. Kiemle, K. Molch, S. Schropp, N. Weiland, and E. Mikusch, "Big data management in Earth observation: the German satellite data archive at the German Aerospace Center," in *IEEE Geoscience and Remote Sensing Magazine*, vol. 4no. 3, pp. 51–58, 2016.

[19] G. Crump, "How to protect a big data archive," *Information Week*, vol. 1, no. 1333, p. 17, 2012.

[20] M. Kandidayeni, H. Chaoui, L. Boulon, and J. P. Trovão, "Adaptive Parameter Identification of a Fuel Cell System for Health-Conscious Energy Management Applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, p. 99, 2021.

[21] P. Liu, Y. Zhang, T. Fu, and J. Hu, "Intelligent Mobile Edge Caching for Popular Contents in Vehicular Cloud Toward 6G," *IEEE Transactions on Vehicular Technology*, vol. 1, no. 1, p. 99, 2021.

[22] G. M. Hall and J. Howe, "Energy from waste and the food processing industry," *Process Safety and Environmental Protection*, vol. 90, no. 3, pp. 203–212, 2012.

[23] J. Wan, J. Zhou, and X. Gui, "Intelligent Rack-Level Cooling Management in Data Centers with Active Ventilation Tiles: A Deep Reinforcement Learning," *Intelligent Systems,IEEE*, vol. 1, no. 1, p. 99, 2021.

[24] X. Li, Q. Yu, B. Alzahrani et al., "Data Fusion for Intelligent Crowd Monitoring and Management Systems: A Survey," *IEEE Access*, vol. 1, no. 1, p. 99, 2021.

[25] A. Priyanka, M. Parimala, K. Sudheer, R. Kaluri, K. Lakshmanna, and M. P. K. Reddy, "*BIG data based on healthcare analysis using IOT devices*," *IOP Conference Series: Materials Science and Engineering*, vol. 263, 2017no. 4, Article ID 042059, 2017IOP Publishing, 2017.

[26] C. Menelaou, S. Timotheou, P. Kolios, and C. G.. Panayiotou, "Joint Route Guidance and Demand Management for Real-Time Control of Multi-Regional Traffic Networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 14, p. 99, 2021.

[27] J. P. Ortiz, G. P. Ayabaca, A. R. Cardenas, D. Cabrera, and J. D. Valladolid, "Continual Refoircement Learning Using Real-World Data for Intelligent Prediction of SOC Consumption in Electric Vehicles," *Latin America transactions*, vol. 20, no. 4, pp. 51–58, 2022.

[28] F. Bonofiglio, M. Schumacher, and H. Binder, "Recovery of original individual person data (IPD) inferences from empirical IPD summaries only: applications to distributed computing under disclosure constraints," *Statistics in Medicine*, vol. 39, no. 8, pp. 1183–1198, 2020.

[29] S. Li and S. Avestimehr, "Coded computing: mitigating fundamental bottlenecks in large-scale distributed computing and machine learning," *Foundations and Trends® in Communications and Information Theory*, vol. 17, no. 1, pp. 1–148, 2020.

[30] X. Ding, Q. Shi, B. Cai, T. Liu, Y. Zhao, and Q. Ye, "Learning multi-domain adversarial neural networks for text classification," *IEEE Access*, vol. 7, pp. 40323–40332, 2019.

[31] C. Lin, C. J. Hsu, Y. S. Lou et al., "Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes," *Journal of Medical Internet Research*, vol. 19, no. 11, article e380, 2017.

[32] M. A. Al-Khasawneh, I. Uddin, S. A. A. Shah, A. M. Khasawneh, L. Abualigah, and M. Mahmoud, "An improved chaotic image encryption algorithm using Hadoop-based MapReduce framework for massive remote sensed images in parallel IoT applications," *Cluster Computing*, vol. 25, no. 2, pp. 999–1013, 2022.

[33] N. B. Mahiddin, Z. A. Othman, A. A. Bakar, and N. A. Rahim, "An Interrelated Decision-Making Model for an Intelligent Decision Support System in Healthcare," *IEEE Access*, vol. 10, pp. 31660–31676, 2022.

[34] D. K. Mohanty, A. K. Parida, and S. S. Khuntia, "Financial market prediction under deep learning framework using auto encoder and kernel extreme learning machine," *Applied Soft Computing*, vol. 99, article 106898, 2021.

[35] D. Li, S. Li, S. Zhang, J. Sun, L. Wang, and K. Wang, "Aging state prediction for supercapacitors based on heuristic Kalman filter optimization extreme learning machine," *Energy*, vol. 250, article 123773, 2022.

[36] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.