A Digital Management System for Archive Resources Based on Cloud Storage in the New Media Era

Qi Yang[a], Jinghuan Zhu[b], Shilong Wang [c*]

Guangxi Science & Technology Normal University, Laibin, China

[a]yangqi@gxstnu.edu.cn, [b]zhujinghuan@gxstnu.edu.cn,
[c*]wangshilong@gxstnu.edu.cn
*Corresponding author: wangshilong@gxstnu.edu.cn

Abstract. In order to solve the problem of incomplete information extraction in the existing digital conversion process of archives, a preset area scanning unit is set up, which is used to activate the OCR scanner of the paper scanning device to scan the preset area when the user operation interface receives the archive upload command, and obtain the first OCR scanning file, which has table distribution characteristics and text distribution characteristics. The file template matching unit is used to match templates based on table distribution characteristics and text distribution characteristics, combined with the file template matching algorithm to generate file matching templates. Template file comparison unit, used to match the template with the first OCR scan file based on the archive, obtain missing attribute information and distribution location of missing attributes. Local compensation scanning unit, used to activate the OCR scanner for local compensation scanning based on missing attribute information and the distribution position of missing attributes, and generate a second OCR scanning file. The archive retrieval result unit is used to synchronize the second OCR scanned file, archive matching template, and project ID information to the cloud server. Based on the project ID information and archive matching template, secondary retrieval is performed in the cloud storage repository embedded in the cloud server to generate digital archive retrieval results. The archive classification result unit is used to activate the archive classification module to adjust the timing of the second OCR scanning file and digital archive retrieval results, generate archive classification results, and update the cloud storage repository.

Keywords: Cloud storage; Digitization of archives; Management system

1 Introduction

Cloud storage is a model of online storage that stores data on multiple virtual servers typically hosted by third parties, rather than on dedicated servers [1-3]. Hosting companies operate large data centers, and those who need data storage hosting can meet their data storage needs by purchasing or renting storage space from them. Data center operators prepare storage virtualization resources on the backend based on customer

B. K. Kandel et al. (eds.), *Proceedings of the 2023 3rd International Conference on Business Administration and Data Science (BADS 2023)*, Atlantis Highlights in Computer Sciences 19,
https://doi.org/10.2991/978-94-6463-326-9_33

needs, and provide them in the form of storage resource pools. Customers can then use this storage resource pool to store files or objects on their own [4-6]. With the rapid development of computer technology, communication technology, and internet technology, traditional archive management models have also encountered severe challenges. Compared with developed countries, there is still a lag in the modernization of archive management in China. In the past, extensive models have been used to solve this problem by increasing office staff and expenses, resulting in a significant increase in management costs. And digital management of archives transforms traditional paper-based archival information objects into machine readable archives, which not only saves storage costs and space, but also makes it extremely convenient and quick to access, thus avoiding the waste of paper and personnel caused by repeated printing of materials [7-10]. We should fully utilize modern technology to transform traditional archive management methods, accelerate the construction of electronic archives, improve the cadre archive management system and cadre information management system, and gradually achieve digitalization of archive management [11]. Digitalization of archives will inevitably become the main form of existence for archives in the future [12].

2 Design ideas

As shown in Figure 1, Firstly, at the level of digital transformation, a technical solution combining template matching and compensation scanning is adopted. Through a series of processing processes such as initial OCR recognition of preset areas, comparison with templates to obtain missing information, and compensation scanning of missing parts, a complete transformation of the content of paper archives is achieved, effectively solving the problem of incomplete information extraction and improving the quality of digital transformation.

Secondly, at the level of digital archive storage, cloud storage technology is adopted to upload complete digital archives to a unified cloud storage repository, achieving centralized management of archives, facilitating unified query, access, and utilization of archives, and solving the problem of distributed storage of digital archives. Finally, at the level of digital archive processing, design an archive classification module that can automatically classify and process digital archives based on their content, intelligently organizing and processing them to meet subsequent management needs.

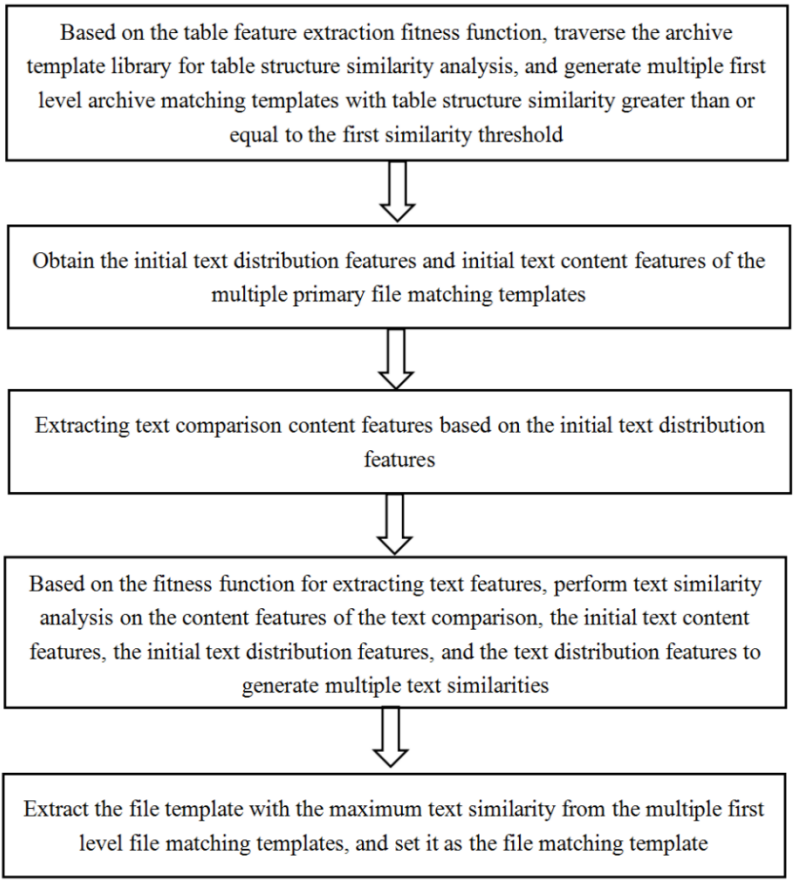

Fig. 1. Technical design roadmap

3 Design Cases

As shown in Figure 2, this case provides a digital archive management system based on cloud storage, which includes a cloud server and a user end, a paper scanning device and a user operation interface, a cloud server including an archive classification module, and a preset area scanning unit, Used to activate the OCR scanner of the paper scanning device to scan the preset area and obtain the first OCR scanning file when the user interface receives the file upload command. The first OCR scanning file has table distribution features and text distribution features. The file template matching unit is used to match templates based on table distribution characteristics and text distribution characteristics, combined with the file template matching algorithm to generate file matching templates. The template file comparison unit is used to compare the file matching template with the first OCR scan file to obtain missing attribute information and distribution location of missing attributes. The local compensation scanning unit is

used to activate the OCR scanner for local compensation scanning based on missing attribute information and the distribution position of missing attributes, and generate a second OCR scanning file. The archive retrieval result unit is used to synchronize the second OCR scanned file, archive matching template, and project ID information to the cloud server. Based on the project ID information and archive matching template, secondary retrieval is performed in the cloud storage repository embedded in the cloud server to generate digital archive retrieval results. The archive classification result unit is used to activate the archive classification module to adjust the timing of the second OCR scanning file and digital archive retrieval results, generate archive classification results, and update the cloud storage repository.

The preset area scanning unit includes the following execution steps: collecting and scanning PDF sets, horizontal cell quantity identification sets, vertical cell quantity identification sets, and cell four point coordinate identification sets. Collect and scan PDF sets, text content identification sets, and text location identification sets. Supervise the first attention channel of the convolutional neural network structure with a set of horizontal cell quantity identifiers, a set of vertical cell quantity identifiers, and a set of cell four point coordinate identifiers, train the scanned PDF set, and generate table feature extraction nodes. Supervise the second attention channel of convolutional neural network structure with text content identification set and text position identification set, train the scanned PDF set, and generate text feature extraction nodes. Extract table structure features from the first PDF file scanned by the OCR scanner based on the table feature extraction node, and generate table distribution features. Extract text features from the first PDF file based on the text feature extraction node to generate text distribution features. Aggregate and associate table distribution features with text distribution features to generate the first OCR scan file.

The preset area scanning unit also includes the following execution steps:

Construct a table feature extraction fitness function:

$$\mathrm{LOSS}_1 = \begin{cases} \dfrac{w_1}{|n_{i1} - n'_{i1}| + |n_{i2} - n'_{i2}|} + \dfrac{w_2 m}{\sum_{j=1}^{m}(0.25 * \sum_{k=1}^{4} d_{jk})}, & [|n_{i1} - n'_{i1}| \le a] \wedge [|n_{i2} - n'_{i2}| \le a] \\ 0, & [|n_{i1} - n'_{i1}| > a] \vee [|n_{i2} - n'_{i2}| > a] \end{cases}$$

Among them, LOSS_1 representing the output fitness of any set of table feature extraction training, n_{i1} representing the number of horizontal cells extracted from the i-th group of training data, n'_{i1} representing the number of horizontal cells identified from the i-th group of training data, n_{i2} representing the number of vertical cells extracted from the i-th group of training data, n'_{i2} representing the number of vertical cells identified from the i-th group of training data, w_1 are the first preset weights, d_{jk} representing euclidean distance between the k-th coordinate output value and the identification value of the four point coordinates of the j-th cell, m represents the total number of cells, and a represents the fitness calculation threshold.

Construct a fitness function for text feature extraction:

$$\mathrm{LOSS}_2 = \begin{cases} \dfrac{1}{Q}\sum_{P=1}^{Q}\left(\dfrac{w_3}{L_P} + \dfrac{w_4}{S_P}\right), & (L_P \le b) \wedge (S_P \le c) \\ 0, & (L_P > b) \vee (S_P > c) \end{cases}$$

Among them, $LOSS_2$ represents the output fitness of any set of text feature extraction training, L_P represents the output text content of the p-th cell and the Hamming distance of the text content identification, S_P represents the output text position of the p-th cell and the Euclidean distance of the text position identification, w_3 represents the third preset weight, w_4 represents the fourth preset weight, Q represents the total number of cells with text, b represents the Hamming distance threshold, and c represents the text distribution distance threshold.

Train the table feature extraction node and text feature extraction node based on the table feature extraction fitness function and text feature extraction fitness function. The file template matching unit includes the following execution steps: extracting fitness functions based on table features, traversing the file template library for table structure similarity analysis, and generating multiple first level file matching templates with table structure similarity greater than or equal to the first similarity threshold. Obtain the initial text distribution features and initial text content features of multiple first level file matching templates. Extracting text distribution features based on initial text distribution features for text comparison content features. Based on the fitness function of text feature extraction, text similarity analysis is performed on text comparison content features, initial text content features, initial text distribution features, and text distribution features to generate multiple text similarities. Extract the file template with the maximum text similarity from multiple first level file matching templates and set it as the file matching template.

The file template matching unit also includes the following execution steps: when the number of file templates with the maximum text similarity is 1, set it as the file matching template. When the number of file templates with the maximum text similarity is not 1, extract file templates with text similarity greater than or equal to the text similarity threshold, and send them to the user operation interface on the user end to obtain user feedback information, including file matching templates.

The archive retrieval result unit includes the following execution steps: performing a first level retrieval in the cloud repository based on the project ID information to obtain the first level digital archive retrieval results. Perform a secondary search on the first level digital archive search results based on the archive matching template, obtain the digital archive search results, including: when the number of archives in the digital archive search results is 0, create a new archive template ID in the sub menu of project ID information, store the second OCR scanned file in the new archive template ID, and update the cloud storage repository. When the number of files in the digital archive search result is not 0, activate the archive classification module to execute the classification process.

In summary, any step of the method described above can be stored as computer instructions or programs in unrestricted computer memory and can be recognized by unrestricted computer processors. The first or second may not only represent order relationships, but may also represent a specific concept, and/or refer to multiple elements that can be individually or completely selected.

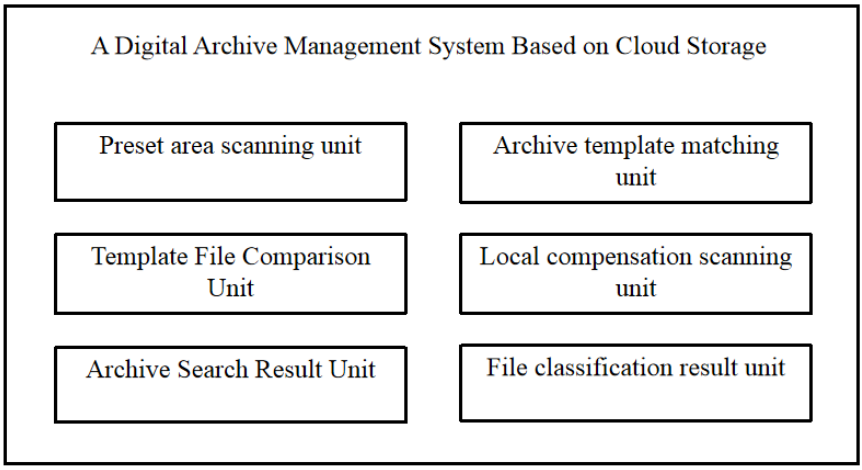

Fig. 2. Module diagram of management system composition

4 Conclusion

Due to the use of an activated OCR scanner to scan the preset area and obtain the first OCR scanned file, the initial recognition of archive content is achieved, improving information extraction efficiency. Generate a file matching template based on table and text features, combined with file template matching, to use template information for comparison and avoid omissions caused by direct full-text recognition. Compare the file matching template with the first OCR scanned file to obtain the distribution of missing information, identify the differences between the scanned file and the complete template, and locate the specific location of missing information. Perform compensation scanning based on missing information to obtain a second OCR scanning file, thereby automatically compensating for any missing issues and ensuring the integrity of the information. Synchronize scanned files to the cloud server for secondary retrieval to generate digital archive retrieval results, achieving centralized management of archives. The archive classification module adjusts the timing, generates archive classification results, and updates the cloud storage repository to achieve intelligent classification of digital archives, facilitate subsequent management, and complete the technical solution for digital preservation of archives. This solves the technical problem of incomplete information extraction and lack of effective compensation in the existing archive digital conversion process, resulting in poor archive digital management effectiveness, and improves the integrity of archive information extraction Implement automatic compensation to enhance the technical effectiveness of digital archive management.

Acknowledgments

Qi Yang was the first author, Shilong Wang was the corresponding author. Guangxi Science & Technology Normal University was the first author's unit. This work were financially supported by Research Project on the Theory and Practice of Ideological and Political Education for College Students in Guangxi in 2021 (2021SZ211) and 2022 Guangxi University Middle-aged and Young Teachers' basic Scientific Research Ability Improvement Project (2022KY0844) and 2022 Philosophy and Social Sciences Research Project in Laibin City (2022LBZS015) and The Second Batch of Industry University Cooperation Collaborative Education Projects of the Ministry of Education in 2021 (202102015063).

References

1. Shixu Zhang, Yaowang Li, Ershun Du, et al. A review and outlook on cloud energy storage: An aggregated and shared utilizing method of energy storage system[J], Renewable and Sustainable Energy Reviews, 2023, 185, 113606.
2. Tong Zhang, Wei Zhang, Jiahui Wang, et al. Design and Implementation of Massive Key Management Scheme for Cloud Storage [J]. Information Security Research, 2023, 9 (9): 859-867.
3. Liya Di. Research on the Management Process of Digital Archives Resources in the Cloud Environment [J]. Archives Research, 2014 (5): 71-75.
4. HAG X, CHEN H, JIAC F, et al. A Secure Deduplication Scheme Based on Data Popularity with Fully Random Tags, 2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2021, 207-214.
5. Chenam Venkata Bhikshapathi, Ali Syed Taqi. A certificateless authenticated searchable encryption with dynamic multi-receiver for cloud storage [J]. Computer Communications, 2023, 211:157-177.
6. Yongjun Xu, Zhen Zhang, Qionghui Ren. The functional relationship between archive departments and data management departments in the context of the national big data strategy [J]. Library and Information Work, 2019 (18): 5-13.
7. Zongsheng Cui. Research on Digital Processing Management of Archives Based on Aerospace Security and Confidentiality Experience [J]. Archives of China, 2023 (8): 42-44.
8. Xiang He. Discussion on Digital Management of Archives and Information Security [J]. Lan Tai Nei Wai, 2023 (2): 4-6.
9. Ahmad Shahnawaz, Mehfuz Shabana. Efficient time-oriented latency-based secure data encryption for cloud storage [J]. Cyber Security and Applications, 2024, 2, 100027.
10. Chaudhary Ajay, Peddoju Sateesh K, Chouhan Vikas. Secure Authentication and Reliable Cloud Storage Scheme for IoT-Edge-Cloud Integration [J]. Journal of Grid Computing, 2023, 21, 3.
11. Yanhua Cao. Analysis of Informationization in Library Archives Management in the Digital Era [J]. Archives Management, 2019 (4): 95-96.
12. Wang Fang. Exploration of Hospital Archives Management in the Context of Big Data [J]. Archives and Construction, 2017 (11) :92-93,84.