



Approaches and Strategies for Digital Construction of Archive Resources Based on Cloud Storage

Qi Yang^a, Jinghuan Zhu^b, Shilong Wang^{c*}

Guangxi Science & Technology Normal University, Laibin, China


^ayangqi@gxstnu.edu.cn, ^bzhujinghuan@gxstnu.edu.cn,
^c*wangshilong@gxstnu.edu.cn

*Corresponding author: wangshilong@gxstnu.edu.cn

Abstract. In order to solve the technical problem of incomplete information extraction and lack of effective compensation in the digital transformation process of existing archives, which leads to poor digital management of archives, the preset area is scanned to obtain the first OCR scanned file. Perform template matching and generate file matching templates. Compare the file matching template with the first OCR scan file to obtain missing attribute information and the distribution location of missing attributes. Perform local compensation scanning to generate a second OCR scanning file. Conduct secondary retrieval to generate digital archive retrieval results. Adjust the timing of the second OCR scan file and digital archive retrieval results, generate archive classification results, and update the cloud storage repository. It can achieve the technical effect of improving the integrity of archive information extraction, achieving automatic compensation, and thus enhancing the effectiveness of archive digital management.

Keywords: Cloud storage; Archives; Digitization; Administration

1 Introduction

Cloud storage is a new concept that extends and derives from the concept of cloud computing . Cloud computing is the development of distributed computing, parallel computing, and grid computing. It involves automatically dividing a large computing program into countless smaller subprograms through the network, and then handing them over to a large system composed of multiple servers for calculation and analysis, and then transmitting the processing results back to users [4-6]. Through cloud computing technology, network service providers can process tens or even billions of information in seconds, achieving network services as powerful as "super-computers" [7-9]. With the development of information technology, digital management of archives has become an important part of improving work efficiency and promoting information construction. In the current information extraction process of archive digitization, information recognition is mainly achieved by scanning the entire archive or manually inputting it page by page. However, due to the complexity of file

layout and content, direct scanning and recognition of the entire file often cannot achieve good results. And when there is a lack of information, there is a lack of effective remedial measures to supplement complete digital archive information [10-11].

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.

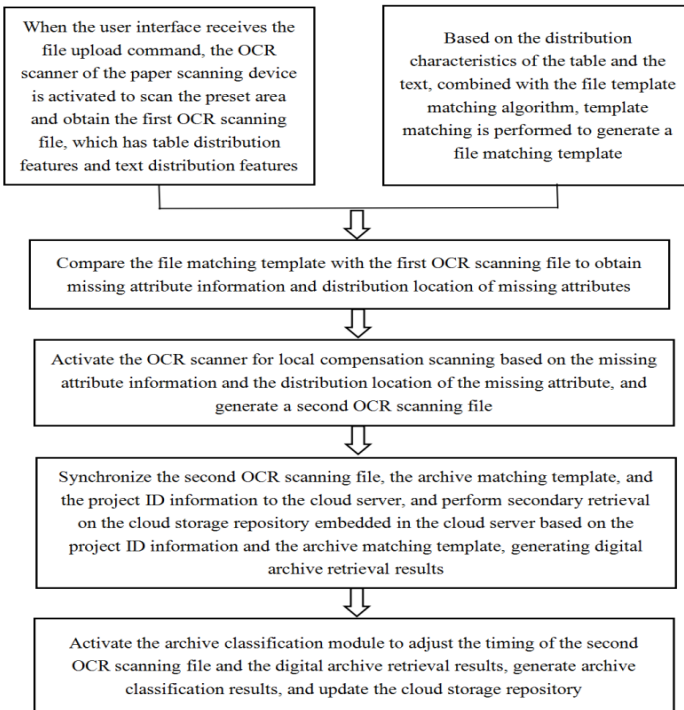


Fig. 1. Technology construction approach and strategy diagram

3 Design Cases

When the user interface receives the file upload command, the OCR scanner of the paper scanning device is activated to scan the preset area and obtain the first OCR scanning file, which has table distribution features and text distribution features.

This step specifically includes: first, collecting and scanning PDF sets, horizontal cell quantity identification sets, vertical cell quantity identification sets, and cell four point coordinate identification sets. The horizontal cell quantity identification set, vertical cell quantity identification set, and cell four point coordinate identification set supervise the first attention channel of the convolutional neural network structure, train the scanned PDF set, and generate table feature extraction nodes. The text content identification set and text position identification set supervise the second attention channel of the convolutional neural network structure, 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.

In a preferred case, a table feature extraction node and a text feature extraction node are first constructed to obtain the table distribution features and text distribution features, thereby obtaining the first OCR scan file. Firstly, a scanner is used to scan paper forms and generate a set of electronic PDF files containing different forms as the scanned PDF set. Manually annotate the number of horizontal and vertical cells in each PDF file, and record them as a set of horizontal and vertical cell quantity identifiers. Using manual framing, label the four point coordinates of each cell in each table to obtain a set of cell four point coordinate identifiers. At the same time, identify and scan the text content in each PDF file table in the PDF set, generate a text string, and obtain the text content identification set. Determine the precise coordinate position of each text content in the document to obtain a set of text position identifiers.

Then, a convolutional neural network model is constructed, which includes the first attention channel for table feature extraction. Supervise the training of the first attention channel using the collected horizontal cell quantity identification set, vertical cell quantity identification set, and cell four point coordinate identification set. During the training process, continuously optimize the ability of the first attention channel to extract table features by scanning PDF documents, so that its output results match the identification set. When the first attention channel can accurately extract table features, the first attention channel is the generated table feature extraction node for training. Meanwhile, the convolutional neural network model includes a second attention channel for text feature extraction. Supervise the training of the second attention channel using the collected text content identification set and text position identification set. During the training process, continuously optimize the ability of the second

attention channel to extract text features by scanning PDF documents, so that its output results match the identification set. When the second attention channel can accurately extract text content and position features, the second attention channel is the text feature extraction node generated by training.

When the user issues a file upload command through the operating interface, the system automatically activates the OCR scanner in the paper scanning device to scan the preset area. Among them, the preset area is the scanning and placement area for paper files, and the files placed in the preset area to be scanned are the first PDF files. Then, use the trained table feature extraction nodes to analyze the table structure of the first PDF file, and obtain the table distribution features, such as the number of rows, columns, and cells. Use text feature extraction nodes to analyze the text content and position of the first PDF file, and obtain text distribution features.

Thirdly, using the distribution characteristics of tables, determine the coordinate range of each table cell in the document image. Using the distribution characteristics of text, determine the coordinate position of each text content. Traverse all text content to determine if its coordinates fall within the coordinate range of the table cells. For text located within a table cell, establish an association between the text content and the corresponding cell, organize it into an electronic document format, and generate the first OCR scan file.

This step also includes constructing a table feature extraction fitness function:

$$LOSS_1 = \begin{cases} \frac{w_1}{|n_{i1} - n'_{i1}| + |n_{i2} - n'_{i2}|} + \frac{w_2 m}{\sum_{j=1}^m (0.25 * \sum_{k=1}^4 d_{jk})}, & [|n_{i1} - n'_{i1}| \leq a] \wedge [|n_{i2} - n'_{i2}| \leq a] \\ 0, & [|n_{i1} - n'_{i1}| > a] \vee [|n_{i2} - n'_{i2}| > a] \end{cases}$$

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, The Euclidean distance between the k -th coordinate output value and the identification value d_{jk} representing 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:

$$LOSS_2 = \begin{cases} \frac{1}{Q} \sum_{p=1}^Q \left(\frac{w_3}{L_p} + \frac{w_4}{S_p} \right), & (L_p \leq b) \wedge (S_p \leq c) \\ 0, & (L_p > b) \vee (S_p > c) \end{cases}$$

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.

In a preferred case, in order to improve the accuracy of table feature extraction nodes and text feature extraction nodes in extracting PDF files, the table feature extraction fitness function and text feature extraction fitness function are preferred.

Among them, the fitness function for table feature extraction is:

$$LOSS_1 = \begin{cases} \frac{w_1}{|n_{i1} - n'_{i1}| + |n_{i2} - n'_{i2}|} + \frac{w_2 m}{\sum_{j=1}^m (0.25 * \sum_{k=1}^4 d_{jk})}, & [|n_{i1} - n'_{i1}| \leq a] \wedge [|n_{i2} - n'_{i2}| \leq a] \\ 0, & [|n_{i1} - n'_{i1}| > a] \vee [|n_{i2} - n'_{i2}| > a] \end{cases}$$

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, The Euclidean distance between the k -th coordinate output value and the identification value d_{jk} representing the four point coordinates of the j -th cell, m represents the total number of cells, and a represents the fitness calculation threshold.

Quantify the training effect of table feature extraction through the fitness function $LOSS_1$ of table feature extraction, characterize the quality of training output, and reflect the similarity between the table features extracted by the first attention channel and the real table features. Compare n_{i1} and n'_{i1} will be conducted, Compare n_{i2} and n'_{i2} at the same time, if the difference between the two is less than or equal to the fitness calculation threshold a , then the table feature extraction fitness will be performed $LOSS_1 = \frac{w_1}{|n_{i1} - n'_{i1}| + |n_{i2} - n'_{i2}|} + \frac{w_2 m}{\sum_{j=1}^m (0.25 * \sum_{k=1}^4 d_{jk})}$. When n_{i1} and n'_{i1} deviation or n_{i2} and n'_{i2} deviation from is greater than the fitness calculation threshold a , the fitness of the table feature extraction is $LOSS_1 = 0$. Among them, the higher the fitness function $LOSS_1$ for table feature extraction, the closer the output result of the channel is to the identification value, and the more accurate the extraction result is.

Among them, the fitness function for text feature extraction is:

$$LOSS_2 = \begin{cases} \frac{1}{Q} \sum_{p=1}^Q \left(\frac{w_3}{L_p} + \frac{w_4}{S_p} \right), & (L_p \leq b) \wedge (S_p \leq c) \\ 0, & (L_p > b) \vee (S_p > c) \end{cases}$$

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.

Quantify the training effect of text feature extraction through the fitness function $LOSS_2$ of text feature extraction, characterize the quality of training output, and reflect the similarity between the text features extracted by the second attention channel and the real text features. If the Hamming distance L_p between the text content and the text content identification is less than or equal to the Hamming distance threshold b , and the Euclidean distance S_p between the output text position and the text position identification is less than or equal to the text distribution distance threshold c , then $LOSS_2 = \frac{1}{Q} \sum_{P=1}^Q \left(\frac{w_3}{L_p} + \frac{w_4}{S_p} \right)$. On the contrary, then $LOSS_2 = 0$. Among them, the higher the fitness function $LOSS_2$ for text feature extraction, the closer the text content and position distribution output by the channel are to the identification value, and the more accurate the extraction results are.

Then, based on the constructed table feature extraction fitness function and text feature extraction fitness function, train the table feature extraction node and text feature extraction node to accurately extract the table and text features of the archive. Firstly, prepare to scan the PDF set as the training dataset, which annotates the number of rows and columns, cell coordinates, etc. of the table as the identification of table features, and also annotates the text content and position as the identification of text features. For each training sample, the extraction results of table and text features are obtained through forward calculation through table feature extraction nodes and text feature extraction nodes. Based on the extraction results and identification, calculate the fitness function $LOSS_1$ and $LOSS_2$ for table feature extraction and text feature extraction. Using a backpropagation algorithm, use the fitness function as the loss function to update network parameters to minimize the fitness function and continuously approach the identification target. Iterating the above process until the network converges, finally obtaining the trained table feature extraction node and text feature extraction node. Based on the distribution characteristics of tables and text, combined with the archive template matching algorithm, template matching is performed to generate archive matching templates.

4 Conclusion

When the user interface receives the file upload command, the OCR scanner of the paper scanning device is activated to scan the preset area and obtain the first OCR scanning file. The first OCR scanning file has table distribution features and text distribution features, thereby quickly obtaining the initial digital content of the file. Based on the distribution characteristics of tables and text, combined with the archive template matching algorithm, template matching is performed to generate archive matching templates, providing a reference basis for locating and extracting accurate information from archives. Compare the file matching template with the first OCR scanned file to obtain missing attribute information and the distribution location of missing attributes, and achieve accurate identification of missing information locations. Based on the missing attribute information and the distribution location of missing attributes,

activate the OCR scanner for local compensation scanning, generate a second OCR scanning file, and scan and complete the initial missing information in a targeted manner to achieve automatic compensation and ensure the integrity of the file. Synchronize the second OCR scan file, archive matching template, and project ID information to the cloud server. Based on the project ID information and archive matching template, perform secondary retrieval in the cloud storage repository embedded in the cloud server, generate digital archive retrieval results, achieve centralized management of digital archive data, and improve the efficiency of digital archive query. Activate the archive classification module to adjust the timing of the second OCR scanning file and digital archive retrieval results, generate archive classification results, update the cloud storage repository, and achieve intelligent classification of digital archives, making their organizational structure clearer, management more convenient, and improving 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. 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.
2. Bhadauria R, Chakia R, Chaki N, et al. A survey on securityissues in cloud computing [J]. *arXiv preprint, arXiv: 1109.5399*, 2011.
3. CHHABRA N, BALA M. An Optimized Data Duplication Strategy for Cloud Computing: Dedup with ABE and Bloom Filters[J]. *International Journal of Future Generation Communication and Networking*, 2020, 13(1): 824-834.
4. Yingmei Li. Research on the Construction Path of Archive Information Resource Sharing Cloud System [J]. *Heilongjiang Archives*, 2023 (3): 22-24.
5. Xiaowen Li. Application of Cloud Technology in Hospital Archives Management and Design Concept of Service Platform [J]. *Office Business*, 2022 (11): 184-186.
6. MAO Xiangjie, ZHANG Pin. Hybrid Verification Scheme for Data Integrity of Cloud Platform[J]. *Computer Engineering*,2020,46(10):46-51.
7. Hongsheng Zhao. Research on the Security Backup Strategy of Digital Archive Resources [J]. *Lantai World*, 2022 (4): 112-114.

8. XIAN Hequn, LIU Hongyan, ZHANG Shuguang, et al. Verifiable Secure Data Deduplication Method in Cloud Storage[J]. Journal of Software,2020,31(2):455-470.
9. Man Han. Exploring the Centralized Management Mode of Electronic Archives in Universities Based on Cloud Computing [J]. Education Practice, 2019 (8): 125-126.
10. Huaying Lian. The Application of Cloud Storage Technology in Electronic Archive Management [J]. Lan Tai Wai Wai, 2022 (8): 19-21.
11. Yuenan Liu. Data Governance: A New Perspective and Function of Archive Management in the Era of Big Data [J]. Archives Research, 2020 (5): 50-57.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

