ISSN 2616-7433 Vol. 6, Issue 8: 84-88, DOI: 10.25236/FSST.2024.060815

Analysis of Unstructured Document Data Extraction Technology

Shiguang Sun[1,a,*], Jiaxin Ye[2,b]

[1]*School of Innovation and Entrepreneurship, Liaoning University, Shenyang, Liaoning, China*
[2]*Asia-Australia Business College, Liaoning University, Shenyang, Liaoning, China*
[a]*sunshiguang@lnu.edu.cn*, [b]*794653291@qq.com*
[*]*Corresponding author*

Abstract: *In real life, a considerable quantity of unstructured documents exists. These unstructured documents are characterized by low structure, high proportion of data storage, abundant information volume and redundant data, which make the management of such data extremely complex. A large amount of valuable data is encompassed within unstructured documents. The key to extract unstructured document information lies in overcoming the difficulties of traditional information extraction techniques, such as, the limitation of model generalization ability, the complexity of context understanding, the difficulty of data annotation, and etc. How to analyze the data of unstructured documents and automatically extract the useful data information therefrom presents challenges to the technologies in the domains of pattern recognition, machine learning, and deep learning. This paper examines the structural characteristics of unstructured documents. It collates and analyzes the existing technologies and methods for extracting information from such documents, with the objective of better serving the digitization of archives.*

Keywords: *Unstructured Document, DLA, Deep Learning, RNN, CNN*

1. Introduction

The digitization of traditional archives has long been underway in academia and industry. After more than ten years of effort, at present, the work of archival digitization covers a wide range, and the amount of data formed is huge. The archives department has carried out digitization work on traditional archives, and the work has achieved results. The proportion of digital archives has increased significantly. Unstructured file data is mostly stored in folders, and picture file is its main file format [1]. These document data picture files are mostly obtained by optical character recognition technology, OCR. How to manage and use these data efficiently, the first task is to complete the conversion of unstructured data to structured data.

As shown in Figure 1, the data extracted from the entire unstructured document is mainly divided into the following parts, including layout analysis (in fact, layout segmentation and area identification), layout logic understanding and layout reconstruction, that is, to reorganize the accurately identified information, so as to obtain editable electronic documents with complete structure.

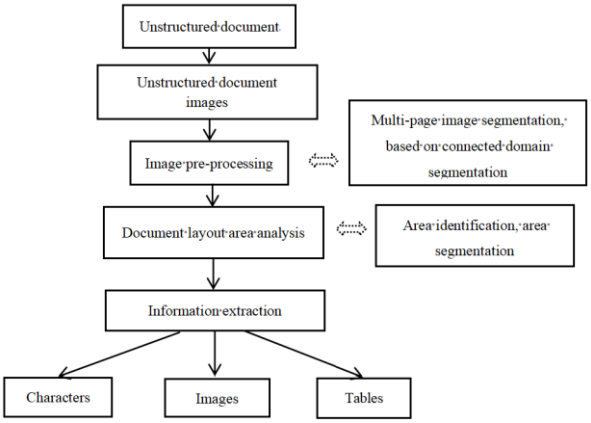

Figure 1: Unstructured document information extraction process

ISSN 2616-7433 Vol. 6, Issue 8: 84-88, DOI: 10.25236/FSST.2024.060815

2. Layout Analysis Technology

Document Layout Analysis (DLA) is a pre-processing step for document understanding. Aims at dividing the document image into different layout areas, such as text, image, table, and etc, and to determine the relationship among these areas. It is a key step in the digitization system and is of great significance for document retrieval, text classification, text recognition, and other tasks. DLA includes pre-processing, layout analysis strategy, post-processing, and performance evaluation, but due to the diversity and complexity of document layout, has not been developed to adapt to all types of document layouts of the general DLA algorithm. With the development of technologies such as deep learning, the accuracy and efficiency of DLA continue to improve, and provide strong support for automated document processing.

The pre-processing of unstructured document image data is to reduce unnecessary noise and establish uniform format and vector representation according to the requirements of subsequent processing algorithms. Without affecting the data quality, reduce the factors that affect the subsequent calculation and processing to ensure the quality of the subsequent processing results [2].

Layout analysis starts with layout segmentation and area identification of the document. The commonly used methods include pattern recognition, region connection, and texture analysis. The main method is to classify samples into a specific category according to the characteristics of samples by learning algorithm according to certain criteria under the condition of known statistical model of research objects or known discriminant function class. The region classification can also be realized by analyzing the texture distribution that reflects different gray levels of pixels and surrounding space [3].

Additionally, real-world documents typically have multiple pages. For this kind of image, the initial step is pagination. Nevertheless, it is challenging to guarantee the optimal outcome of layout segmentation, which will also exert a significant influence on the subsequent segmentation work based on the connected domain. Many existing detection or recognition tasks only recognize and segment single-page images, disregarding the actual two-page documents. To facilitate the processing of double-page images, we often need to undertake a pre-processing to segment the double-page images to form two complete single-page text images. If a single-page image approach is employed to identify and segment the double-page text, numerous errors will arise. The conventional method for double-page images is to manually divide the page and cut the double page into two single pages, but this consumes a considerable amount of time and is relatively inefficient.

To solve this problem, it is essential to consider directly in accordance with the outline of the discovered foreground area, namely, the page outline. Based on the aspect ratio information of the small external matrix outline, determine whether the page is a double-page one, and then perform a paging cut directly on the central axis of the page.

The analysis techniques of unstructured document layout mainly include the following:

(1) Artificial Neural Networks (ANN), we can simply understand, and connect the network with the human neuron structure, the neuron is the node in the network structure, and we compare the neuron trigger response mode to an operation model, When a neuron is activated by receiving a signal and then transmits information to other neurons, the purpose of processing and transmitting information is realized. Based on the geometric coordinates of the text area, visual features, text semantics, and other modal information, predicts the text reading order, and improves the accuracy of classification results.

(2) A method based on deep learning: The deep learning algorithm is used to pre-process the image, such as deformation correction, shadow removal, Mohr stripe removal, and reflection removal, to improve image quality and to provide accurate data basis for subsequent layout analysis. Deep learning models such as convolutional neural networks (CNN) are used to identify and segment text, graph, table, and other elements in the document image, and divide the document image into different content areas. The graph neural network (GNN) and other models are used to analyze the logical relationship between regions, such as reading order, paragraph structure, and etc. enabling the machine to deeply understand the structure and content of the document. Deep learning also plays an important role in table recognition. It can accurately extract the contents and relationships of cells in a table, and can realize automatic processing and analysis of tables. Build a deep learning network to realize the transformation of data from unstructured to structured, such as full convolutional neural network (FCN), improved regional convolutional neural network (FasterR-CNN), and other algorithms. These algorithms based on the semantic segmentation model can analyze the layout accurately and efficiently, and realize the identification, classification, and feature extraction of each part of the layout [4].

ISSN 2616-7433 Vol. 6, Issue 8: 84-88, DOI: 10.25236/FSST.2024.060815

(3) Traditional layout analysis algorithms: The traditional layout analysis method has obvious limitations when dealing with layouts and tables. Is mainly limited by layout differences, and has insufficient generalization effect. These methods usually separate layout segmentation from area recognition, which is computationally heavy and inefficient. In addition, the traditional method is difficult to accurately capture the logical relationship and reading order in the layout, and has limited ability to deal with complex layouts. With the introduction of deep learning technology, these problems have been effectively alleviated. By combining multiple modal information, significantly improves the accuracy and efficiency of layout analysis.

Including top-down, and mixed methods, these methods have important applications in document retrieval, automatic correction, bank document identification, and so on.The top-down approach involves initiating with the entire page or several extensive areas and segmenting the page based on texture features and homogeneity rules. The bottom-up method typically takes pixel points or connected domains as fundamental units, and gradually merges these basic units into large regions in accordance with heuristic rules, and eventually classifies regions based on diverse structural features. The top-down approach is less computationally demanding but can handle less diverse page layouts, while the bottom-up approach can handle more intricate page layouts but is more computationally intensive. The hybrid method that utilizes both methods concurrently can compensate for the deficiencies of the two methods to a certain extent.

(4) Methods based on image processing and pattern recognition: In layout analysis, template matching techniques can be used to locate specific objects or areas in an image. Filter algorithms (such as average filter, median filter, and Gaussian filter) are used to eliminate the noise in the image and to improve the image quality. Useful information in the image is enhanced with techniques such as contrast adjustment, sharpening, and to make the image easier for subsequent processing. Useful information in the image is enhanced with techniques such as contrast adjustment, sharpening, and to make the image easier for subsequent processing. By comparing the predefined template to the areas in the image, one can find the areas that are most similar to the template. Firstly, the unstructured document is converted into a semi-structured image file, and then the pattern recognition method is used to accurately locate each part of the page in the image and finally realize the classification of each part of the page.

3. Information Extraction Technology

Layout information extraction technology is a key step to make the algorithm accurately identify the structure of the document, which mainly includes physical layout analysis and logical layout analysis. Physical layout analysis mainly solves the region segmentation problem, such as region segmentation, classification, text detection and location, text line segmentation, and etc. Furthermore, logical layout analysis focuses on the logical relationship between regions and reading order, such as regional semantic classification, and etc. In addition, layout analysis also involves distinguishing handwriting and printing, form analysis (cell extraction and relationship analysis), and extraction of signature, icon, seal, and other layout elements.

3.1. Text Classification Methods of Cyclic and Convolutional Neural Networks

Recurrent neural network (RNN): RNN is a kind of recurrent neural network that takes sequence data as input, recursively in the direction of sequence evolution, and all nodes are connected by a chain. RNN has the memory function and can handle the text sequence of indefinite length, and solves the text classification problem by capturing the time sequence information [5].

Convolutional neural networks (CNN): CNN is mainly used to extract local features in text classification. It can extract the local features of the text through convolution, and reduce the feature dimension through pooling to improve the classification efficiency. Combining the mixed model of RNN and CNN, such as BGRUCNN, can make use of the ability of RNN to process sequence information and CNN to extract local features, and then better perform text classification.

Before data extraction, text classification methods based on cyclic and convolutional neural networks are adopted to classify unstructured form documents and obtain required form documents, thereby narrowing the scope of subsequent data extraction and improving extraction efficiency and precision. The automatic data extraction model based on deep learning is used to obtain the intermediate semantic vector through bidirectional recurrent neural network coding, and then decrypt

ISSN 2616-7433 Vol. 6, Issue 8: 84-88, DOI: 10.25236/FSST.2024.060815

the intermediate semantic vector through the attention model and single recurrent neural network to obtain the unstructured tabular document data.

Recurrent and Convolutional neural networks (BGRUCNN) are one of the deep learning techniques to extract data information from unstructured documents efficiently. Using the text classification method based on BGRUCNN,the target document is classified first to reduce the scope of extracted information, so as to improve the efficiency of subsequent processing and intensive reading [5].

3.2. Automatic Data Extraction Model based on Deep Learning

Automatic document data extraction technology based on deep learning is an efficient and intelligent data processing method. Building a deep learning model can automatically extract key information from complex document data, such as entities, relationships, events, and etc. The deep learning model has powerful feature extraction and pattern recognition capabilities, can process large-scale and diversified document data, and adapt to different application scenarios [6].

Aiming at the flexible layout structure and complex background, algorithms adopt multi-scale feature extraction technology. By constructing the feature pyramid or using the attention mechanism these algorithms can improve the segmentation accuracy of the network for different size regions, thereby enhancing the overall performance. Optimization algorithms such as gradient descent and backpropagation play an important role in deep learning model training. These algorithms minimize the loss function by constantly updating the parameters of the model, thereby improving the prediction accuracy and generalization ability of the model.

Deep learning models can automatically learn more abstract and high-level feature representations from raw data. The detailed steps include data preparation, feature extraction, model selection and training, model evaluation, and model deployment and application. In the information extraction task, these models are especially good at dealing with such tasks as named entity recognition, and entity relation extraction, which can extract key information from unstructured or semi-structured data, and save it for structured data form. For automation construction of the knowledge base, analysis, text information retrieval application has important significance [7].

For classified unstructured documents, deep learning methods can be used to extract data information. The semantic vector is obtained by bidirectional recurrent neural network BRNN and decrypted by RNN to extract data information from unstructured documents [8].

4. Conclusion

The commonly used method is to analyze the layout contents of documents and pictures by means of projection transformation in an image algorithm. However, such methods usually rely on other complex algorithms and analyze the layout of documents by means of cumbersome pre-processing and morphological processing of pictures, which is relatively time-consuming and cumbersome in the algorithm. The other commonly used method is deep learning. Although this method alone saves the time of layout analysis of document images to a large extent, it has more trouble in the early processing process and requires a large amount of data and a long training time. Under the current situation, the qualified data sets are relatively small, and the space for choice is limited.

The heuristic design algorithm based on connected domain analysis can achieve good results for documents, but these algorithms need manual operation in the process of feature extraction, which is difficult to adapt to a variety of different scenarios. The method based on deep learning is the mainstream method at present, which can automatically extract features and is more general without rule design. Recently, some studies have proposed a multimodal depth model that integrates document images and text. This method based on image fusion can capture the complementary information between images, which is a potential research direction for the document, a medium that naturally contains these two modes.

Unstructured documents are characterized by complex structures and diverse styles. Layout analysis and data extraction are the keys to the digitization of unstructured documents. According to the title information in the unstructured document, the layout is first cut, and then the logical semantic analysis is carried out. Finally, the data of single value range and multi-value range in the unstructured document are extracted comprehensively. Give full play to the advantages of deep learning technology to improve the efficiency and accuracy of unstructured document data extraction.

ISSN 2616-7433 Vol. 6, Issue 8: 84-88, DOI: 10.25236/FSST.2024.060815

Acknowledgments

This work was financially supported by the Educational Reform Project of Liaoning University and the Archives Science and Technology Project of Liaoning Province (Grant No.2023-X-061).

References

[1] Wang Zhiyu, Zhao Shumei. Unstructured data management electronic file analysis. Journal of Archival Science, 2014 (5): 5. DOI: CNKI: SUN: DAXT. 0.2014-05-015.
[2] Zhang Haoyue. Layout Analysis and Table Extraction of Unstructured Documents. Beijing Jiaotong University, 2019.
[3] Zhang Pengfei. Research on Layout Analysis Algorithm of Unstructured Documents. University of Electronic Science and Technology of China [2024-08-21].
[4] Liang Yande. Research on task interference prediction method of Cloud Data Center based on massive log. Beijing University of Technology, 2021.
[5] Xi Jianfei, Wang Zhiying, Zou Wenjing, et al. Unstructured tabular document data extraction method based on Deep learning. Microcomputer Applications, 2022, 38 (2): 4.
[6] Zhang Yunzhen and Tang Wei. Document structural transformation in the process of data processing research. Journal of electronic and information technology, 2021, 005 (002): P. 186-187. The DOI: 10.19772 / j.carol carroll nki. 2096-4455.2021.2.082.
[7] Li Yixin, Zou Yajun, Ma Quanwen. Based on feature extraction and document block image classification algorithm of machine learning. Journal of signal processing, 2019 (5): 11. DOI: 10.16798 / j.i SSN. 1003-0530.2019.05.003.
[8] Gao Yang, Huang Heyan, Lu Chi. A document topic vector extraction method based on deep learning: CN201810748564.1.CN108984526A [2024-08-21].