Volume 3, Issue 12, December 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on the Methods Used in Document Digitization and its Applications

Greeshmamol Varghese*
PG Scholar
Dept. of Computer Science and Engg
Karunya University, India

Kumudha Raimond
Professor
Dept. of Computer Science and Engg
Karunya University, India

Abstract— Document digitization and Document Analysis and Recognition (DAR) are techniques that are used for handling document images. Several techniques were implemented to perform document digitization. Article reconstruction is one of the main applications of document digitization. The four major steps of article reconstruction are grouping the article bodies, detecting the reading order, title body pair association and article parts linking scattered in different pages. This paper presents a survey on different techniques that are used for document digitization as well as for article reconstruction.

Keywords— Document digitization, Document analysis and recognition, Article reconstruction, Data mining, Pattern recognition, Clustering

I. INTRODUCTION

Data mining is an area in computer science that deals with many other sub areas such as pattern recognition, machine learning, artificial intelligence etc. It deals with the raw data as well as the data stored in large database or data warehouse. So in essence in data mining, the stored data will be extracted and processed according to the user requirements. Machine learning and pattern recognition are two evolving areas of application in data mining. The basics of machine learning are classification and text categorization. Pattern recognition includes classification, regression, parsing, labelling, etc. In other words, pattern recognition or pattern matching can be described as a subpart of machine learning. Document digitization is one field in pattern recognition where lots of researches are being carried out. The main application areas of the document digitization are electronic newspaper, information retrieval, printing on demand, digital library and text to speech [1]. There is no common technique for the implementation of these applications. From late 80's onwards various document digitization techniques have been used in the above applications. The applications of the document digitization are not converged to a single field. For each application there will be many steps to be implemented. In those steps the technologies used for document digitization will be different. For finding an efficient and useful technology with suitable methods for document digitization and its applications, a survey on these technologies and methods will be helpful.

The main technologies in the document digitization are document analysis, page segmentation, article reconstruction, text extraction and document clustering. The applications such as digital library and newspaper reconstruction or e-newspaper can be viewed from a single perspective because the technologies and the methods used in both applications are almost similar. Digital library and article reconstruction can be done effectively with the help of four step methodology which is described in an efficient way by Liangcai et al. [1]. In this four step method, the whole technique is divided into grouping the article bodies, detecting the reading order, title body pair association, article parts linking scattered in different pages.

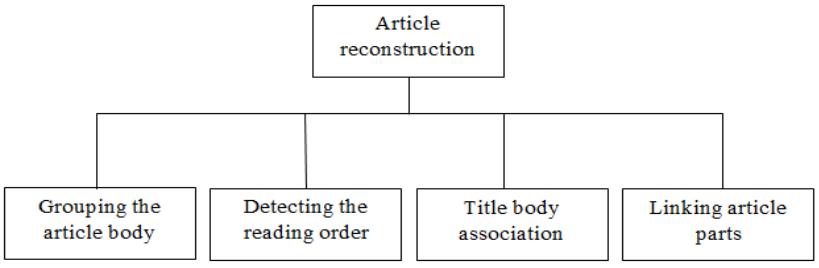

Fig.1 Document Digitization Methodology

In grouping the article bodies, the text blocks are grouped into body blocks and non-body blocks. After grouping, the reading order is detected for the body blocks of each article. Then title corresponding to a single article is combined with

the body blocks. Lastly, article parts scattered in different pages are combined. These four steps are shown in fig. 1. In this paper, the techniques used in the above mentioned four steps for document digitization are analyzed. In this survey, how the digital library and article reconstruction methodologies can be classified and what are the techniques used in each method are tried to solve.

II. SURVEY ON DOCUMENT DIGITIZATION

The existing methods for document digitization can be classified based on

1. Rule based methods
2. Graph or tree based methods
3. Clustering based methods
4. Topological sorting based methods

A. Grouping Based On Rule Based Approach

An integrated system for creating digital library from newspaper archives is proposed by Gatos B. et.al, 1999 [3]. The methodology used in this work is the automatic article clipping using isothetic polygons. Here a rule based approach is used for separation of article body. In this method the output is in image format. The extraction of text from images is not done so the reading order detection is not required in this approach. The clipped article contains both the title and the body so the necessity of the title body association can be avoided in this case. Linking of article parts is done using the link button. Since the addition and subtraction of rectangles is done, no algorithm is used for the methodology. Manual evaluation is used for more accuracy.

Niyogi D. et.al [9] proposed different method for document digitization. This method is named as DELOS Logical derivation system. In this system whitespace analysis algorithm is used for segmentation of articles. Rule based domain knowledge is used for reading order detection and title body association. Pointers are used for linking the article parts. The database used in this system is called Knowledge database. Performance is calculated based on correctly classified original blocks. But the system is complex because of rule based approach.

In automatic page analysis for the creation of digital library [8], another methodology is applied. Here rule based approach is used for segmentation and grouping. Rule based approach is also used for reading order detection of the blocks corresponding to the same article. Run Length Smoothening algorithm (RLSA) algorithm is used for title body association and rule based linking. The performance is calculated using precision and recall. The main advantage of this paper is, it is suitable for complex library digitization.

Phillip E. Mitchell., et. al. [6] in his work proposed an algorithm that performs automated segmentation and classification of images. A feature of the algorithm is a technique for segmenting components that are connected to other components. This means, horizontal lines and vertical lines, which can be useful in determining the page layout, can be segmented from other lines and other components. The algorithm uses a bottom-up approach to initially segment the image, classify the patterns and extract the text lines. The classified patterns are then merged into complete regions. In this technique also RLSA is used. To evaluate the performance, newspaper component detection metric is used. This method is based on counting the number of matches between the regions detected by the algorithm and the regions detected in the ground truth. The major problems for misclassifications are small titles positioned close to text and poor image quality. This method has yielded a success rate of 76%.

B. Grouping Based On Topological Sorting

In [12], two methodologies are suggested for high performance document layout analysis. The methodologies are scale space analysis and appearance based document retrieval. White space analysis algorithm is used for article grouping. Here a topological sorting method is used for reading order detection. But in this paper, title body association is not performed. Another drawback in this paper is that, it can only be used for a single article. So, there is no need for linking the articles separated in different pages. The database used in this case is UW3 (University of Washington 3).

PDF became a very common format for exchanging printable documents. It can be easily generated from the major documents formats, which leads to a huge number of PDF documents' availability over the internet. But its use is limited to displaying and printing, which will considerably reduce the search and retrieval capabilities. Because of this, additional tools have recently appeared that allow extracting the textual content. But their practical use is limited in the sense that the text's reading order is not necessary preserved, especially when handling multi-column documents, or in the presence of complex layout.

K. Hadjar et.al, in [4] proposed a novel approach to overcome the document content extraction. They are by merging 1) low-level extraction methods applied on PDF files with 2) layout analysis performed on a synthetically generated TIFF image. RLSA is used for line merging. This method consists in combining PDF symbol analysis with traditional document image processing techniques. By considering complex layout analysis, it can be results in input to drive a topological sorting algorithm for text pieces. It has brought a significant contribution to improve the reliability of low level text extraction from PDF files.

In [11], topological sorting is used for reading order detection. For article segmentation, Fraunhofer DIU page segmentation module is used. Skew detection algorithm is used in this method. Rule based approach is used for title body association. The performance parameters used are precision and recall. In [10], spatial topology analysis is used for separating the article blocks.

C. Classification Based On Graph Or Tree Based Approach

In hierarchical representation of optically scanned documents [2], a logical representation for form documents to be used for identification and retrieval is devised. A hierarchical structure is proposed to represent the structure of a form by using lines and the XY tree approach. Traditional manual key entering of data on forms is tedious, time-consuming, and error-prone. The main interest in form processing systems is to extract user filled-in data and associate it with the corresponding preprinted field. The reference model form may be stored in a form database when an input instance form is presented to the system. Its type can be identified by matching with one of the reference model forms in the database. Here a heuristic algorithm based on the XY-tree method is described in order to transform the geometric structure of a form document into a hierarchical structure by using horizontal and vertical layout lines which exist on the form. The retrieval of similar forms is performed by computing the edit distances between the generated trees. A cell is defined as the smallest block which only consists of a block frame. However, there is a need for more logical features if the forms have several different physical structures and if the number of forms to be compared is large.

Liangcai Gao, et. al. [1], proposed a graph based method for newspaper article reconstruction. This adopted a bipartite graph method. OCR (Optical Character Recognition) devices are used for segmentation. Optimal matching is used for title body association and reading order detection. For linking, geometric information method is used.

D. Grouping Based On Clustering Method

Non-parametric probability distribution for document representation is used in [13]. A clustering method is used for segmentation of article bodies. Here K-Means clustering algorithm is used and the performance evaluation is done using precision and recall. The main disadvantage of this method is, it is used only for local partition of text blocks, not for global partition.

E. Other approaches

In integrated algorithms for newspaper page decomposition and article tracking [3], a series of applications pertaining to the conversion of newspaper into digital resources as well as transformation of the printed material to an accessible digital achieve are devised. Automatic article clipping using isothetic polygons is used in this paper. A rule based approach is used for segmenting article body. Since the output is an image, no reading order detection is required in this approach. The image will contain both the title and the article body, so the grouping of title and the body blocks is not needed. Linking of blocks in different pages is done with the help of link button. The correctness is measured manually. The margin of the image is decided with the help of isothetic polygons. The main disadvantages of this method are (1) searching method is difficult i.e., the user has to enter additional information to search for a title. (2) User has to scroll the image up and down for reading the newspaper. The segmentation of the newspaper may not be perfect with the help of isothetic polygons. Meta data collected are stored in RDBMS (SQL server 7) and all the low resolution images are stored in file systems.

In [5], studies have been made for document image analysis concerning how to convert the document images into symbolic form to facilitate document storage and retrieval, as well as document modification and reuse. This conversion is a complex process articulated in several steps. After preprocessing, the document image is decomposed into several constituent items which represent coherent components of the documents (e.g., text lines or halftone images. This layout analysis step prepares for the document image understanding, whose aim is that of recognizing semantically relevant layout components (e.g., title and abstract) as well as extracting abstract relationships between layout components (e.g., reading order). WISDOM++ is a document processing system that operates in five steps: document analysis, document classification, document understanding, text recognition with an OCR, and text transformation into HTML/XML format. Some decision tree learning techniques are now applied to the block classification problem that is the separation of text blocks from graphics, while the first release of PLRS used a linear pattern classifier for this task. The symbolic learning technique applied to build the rule base for the document classification and understanding steps has been extended in order to handle both numeric and symbolic data. WISDOM++ can manage multi-page documents, each of which is a sequence of pages.

Ant colony optimization and bipartite graph are used in [1] for newspaper article reconstruction. It considers all the four steps of document digitization. The grouping of article bodies was done with white space analysis algorithm. This method was implemented successfully in many other papers. The main focus of this work is the reading order detection with the help of ant colony optimization. This method is complex but it provides an efficient method for the reading order detection. Title body association and the association of article parts scattered in different pages are done with the help of bipartite graph and optimal matching. For each and every step, performance evaluation has been done based on precision and recall.

III. CONCLUSION

From the survey that has been done based on the four step methodology, each technique uses different methods. It cannot be concluded that only one method can be used for a technique. From the analysis rule based approach did not uses the reading order detection effectively. In graph based method the complexity will be high. Ant colony optimization technique tried to solve the main problems effectively. In this method the reading order detection has is found using pheromone based ant colony optimization. The title body association problem and linking article parts scattered in different pages are tried to solve by optimal matching using bipartite graph. But the system is a little bit complex.

REFERENCES

[1] L. Gaoa, Y. Wanga, Z. Tangb, X. Linc "*Newspaper article reconstruction using ant colony optimization and bipartite graph*" J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] G. Nagy, S. Seth, "*Hierarchical representation of optically scanned documents*", in: Proceedings of the 7th International Conference on Pattern Recognition, IEEE Computer Society, Montreal, Canada, 1984, pp. 347–349.

[3] B. Gatos, S.L. Mantzaris, K.V. Chandrinos, A. Tsigris, S.J. Perantonis, "*Integrated algorithms for newspaper page decomposition and article tracking*", in: Proceedings of the 5th International Conference on Document Analysis and Recognition (Bangalore, India, September 20–22, 1999). ICDAR '99, IEEE Computer Society Press, Los Alamitos, 1999, pp. 559–562.

[4] K. Hadjar, M. Rigamonti, D. Lalanne, R. Ingold, "*Xed: a new tool for extracting hidden structures from electronic documents*", in: Proceedings of 1st International Workshop on Document Image Analysis for Libraries (Palo Alto, CA, USA, January 23–24, 2004). DIAL '04, IEEE Computer Society Press, Los Alamitos, 2004, pp. 212–221.

[5] M. Ceci, A. Appice, C. Loglisci, D. Malerba, "*Preference learning for document image analysis*", in: Proceedings of the ECML/PKDD '10 Tutorial and Workshop on Preference Learning, Barcelona, Spain, September 20–24, 2010.

[6] P. E. Mitchell and Hongan, "*Newspaper Document Analysis featuring Connected Line Segmentation*", School of Electrical and Information Engineering University of Sydney, NSW 2006

[7] M. Aiello, C. Monz, L. Todoran, M. Worring, "*Document understanding for a broad class of documents*", International Journal on Document Analysis and Recognition, 1–16, Apr, 2002.

[8] S.L. Mantzaris, B. Gatos, N. Gouraros, S.J. Perantonis, "*Linking article parts for the creation of a newspaper digital library*", in: Proc. of the Content-Based Multimedia Information Access International Conference, Paris, France, April 2000, 2000, pp. 997–1005.

[9] D. Niyogi, S.N. Srihari, "*Using domain knowledge to derive logical structure of documents"*, in: Proceedings of SPIE, Document Recognition III, San Jose, CA, USA, January 29, 1996, pp. 114–125.

[10] C.H. Papadimitriou, K. Steiglitz, "*Combinatorial Optimization: Algorithms and Complexity*", Prentice Hall, NY, 1982.

[11] F.Y. Shih, S.S. Chen, "*Adaptive document block segmentation and classification*", IEEE Transactions on Systems, Man, and Cybernetics 26 (5) (1996) 797–802

[12] M.B. Thomas, "*High performance document layout analysis*", in: Proceedings of the Symposium on Document Image Understanding Technology, Greenbelt, MD, April 9–11, 2003, SDIUT '03.

[13] R. Huang a, W. Lam, "*An active learning framework for semi-supervised document clustering with language modeling*", Data & Knowledge Engineering 68 (2009) 49–67

