International Journal on New Computer Architectures and Their Applications (IJNCAA) 1(2): 404-411
The Society of Digital Information and Wireless Communications, 2011 (ISSN: 2220-9085)

# Automatic Document Structure Analysis of Structured PDF Files

Rosmayati Mohemad[1, 2], Abdul Razak Hamdan[1], Zulaiha Ali Othman[1] and Noor Maizura Mohamad Noor[2]
[1]Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor Darul Ehsan, Malaysia
[2]Department of Computer Science, Faculty Science and Technology, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu Darul Iman, Malaysia
{rosmayati, maizura}@umt.edu.my, {arh, zao}@ftsm.ukm.my

## ABSTRACT

Portable Document Format (PDF) is the most comfortable way to publish information because of its operating system independent. However, information on PDF document is unstructured and are applicable only for human reader. In addition, PDF consists of non-tagged internal structure which make the extraction task difficult. Automatically details analyzing and recognizing of PDF document structures especially paragraph and tabular area is vital for extracting relevant information precisely for use in other domain applications. Motivation of this study is to support knowledge extraction and exploit its actual semantic for improving further analysis. This paper proposed an intelligent approach to identify and recognize automatically the layout and structure of PDF documents together with their text and then structure the extracted information into ontological-based representation. An experimental study has been conducted using a collection of construction tender documents in PDF to test the performance of the proposed approach. The accuracies of precision, recall and f-measures have shown significant results when detecting tabular and paragraph structure.

## KEYWORDS

Document Analysis, Information Extraction, Portable Document Format.

## 1 INTRODUCTION

Nowadays, government agencies, business corporate companies and personal individual disseminate their electronic documents in Portable Document Format (PDF) as the quickest and convenient way to publish information including text, tables and graphical images. In the survey conducted by Association of Information and Image Management (AIIM) in 2008, 90% of 200 member organizations stored information in PDF either by scanning the documents or converting from Microsoft Office files to PDF-based format and it is predicted the use of PDF fluctuates to 93% for the next five years [1].

Various types of digital documents either newspapers, catalogues, reports, magazines, articles and even forms are available in PDF since the technology is good at offering open standard feature in which it is platform independent for sharing, archiving, retrieving and printing electronic document. Despite of it benefits, PDF has drawback in terms of content and structure analysis. As the result, information represented in PDF format is inconvenient for people to retrieve and reuse in other applications automatically, for example decision-making applications and office automation systems, which requires

machine readable information [2-4]. Organizations are much benefited if they could use and process 80% of their unstructured resources which stored as text [5]. Furthermore, there is lack of information on the internal structure of the PDF document content [6, 7]. Thus, details analyzing and recognizing PDF document structure automatically is vital for extracting and decomposing relevant information both into structure and semantic forms for various purposes such as effective sharing, information searching, decision making and others.

Analyzing and extracting particular text with related structure from PDF document is a non-trivial task due to the existing of various different document layout and structure. A block of text element is defined by its underlying document structure tags such as headings, paragraphs or rectangular boxes (tables) [8]. Normal structure of PDF documents generally may consist paragraph, tabular and image. Zanibbi et al. [9] defined table taxonomy as column, row, cell, block, headers, body and associated text regions. Organized texts either in tabular or paragraph ease human understanding and simplify interpretation. However, machine could not understand the structure of tabular form and related text within the table automatically for further processing. Motivation of this study is to support knowledge extraction and exploit its actual semantic for improving further analysis. This paper proposed an intelligent approach to identify and recognize automatically the layout and structure of PDF documents together with their text and then structure the extracted information into ontological-based representation. The study is carried out in a series of steps to achieve the goal.

The article is structured in the following manner. Literature on past researches and current available tools regarding on PDF document analysis is briefly reviewed in Section 2. Meanwhile, Section 3 presents on the series of steps proposed to analyze the structure and content of PDF document. Next, the experimental setup is provided in Section 4 and Section 5 discusses on the experimental results analysis. Finally, Section 6 concludes with summary of this research.

## 2 RELATED WORKS

PDF document analysis to recognize the fundamental structure of document is the significant research area that received considerable attention from several researchers. Most of them focused on recognizing tabular structure of document. Therefore, different computational approaches have been applied in this research area such as predefined structure models [10], heuristic approach [7, 11] , statistical approach and combination of both heuristic and statistic [12]. In addition, most of this prior researches converting extracted tabular structures and related elements identified on PDF files into other structured format such as HTML and XML format. Hassan [10] proposed wrapper-based approach when detecting table in PDF documents while Jiang and Yang [7] proposed a method to detect PDF document layout and translating it into HTML format. In contrast to our study, recognized text and structure are organized into ontological-based approach which allows for further reasoning process.

Study on this paper was inspired by the research that had been done by Oro and Rufollo [3, 12] where they had proposed

PDF-TREX, a heuristic approach for table recognition and extraction from PDF documents. Here, they carried out analysis on spatial distribution of white spaces between texts for computing horizontal and vertical threshold values to ensure correct text elements were grouped into the same cluster. Threshold value is the fundamental parameter in clustering and the drawback of proposed space distribution analysis is it could cluster dissimilar texts together when the distances among texts were under predetermined threshold. Therefore, we proposed more details analysis on interpreting appropriate distances between texts by identifying the smallest distance measure, thus minimizing in producing wrong cluster. In addition, we proposed rule-based approach when detecting table. The goal of this paper shares the same interest in identifying tabular structure of PDF documents, yet expands to identify paragraph structure and text associated to tabular and paragraph.

There are currently several commercial products available for analyzing PDF documents such as PDF-Analyzer [13] and PDF Analyzer [14]. Schmoekel [13] developed PDF-Analyzer which offers automatic task for retrieving PDF document internal properties, modifying document security level and basic text or content extraction. The tool however does not capable to recognize any paragraph or tabular structure of the document. Meanwhile, PDF Analyzer released by Amyuni Technologies is most similar to the former product in which capable to identify the PDF document internal properties and has been designed to focus more on the consistency checking of the document structure and content defined by some custom rules. Nevertheless, the tool required complicated and skilled techniques to define custom rules and understand the analyzed result. Complicated usage of this tool makes it more suitable for skilled users (developers) than general end user. In addition, the technology also allows for locating and extracting specific text strings, but there is no specific function that able to understand, infer and extract meaningful information in a table, for instance the relation between text in table header and table body.

## 3 AUTOMATIC DOCUMENT ANALYSIS PROCESSES

The fundamental challenge in PDF document analysis is to ensure the approach efficiently recognizes and classifies textual elements associated with or within their appropriate structure and layout especially for tabular structure analysis. The orientation of textual elements and possible paragraph and tabular structures are identified through a series of steps as depicted in Fig. 1. Several steps of this approach are inspired by research that has been done by Oro and Rufollo [12].

### 3.1 Tokenizing

Tokenizing is the process of returning text elements or strings identified on PDF documents into tokens together with their absolute 2-dimensional Cartesian coordinates in the original document. A token is defined as non space characters. The coordinates for each token, $i$ is identified as upper-left ($X_{iL}$, $Y_{iL}$) and bottom-right ($X_{iR}$, $Y_{iR}$) as portrayed in Fig. 2. Here, a token represents a word in the highlighted border. The implementation of this step is accomplished by using Java PDF Extraction Display Access Library

(JPEDAL) that reads and identifies all PDF objects. In addition, it also return the rectangular coordinate of each page of PDF document.

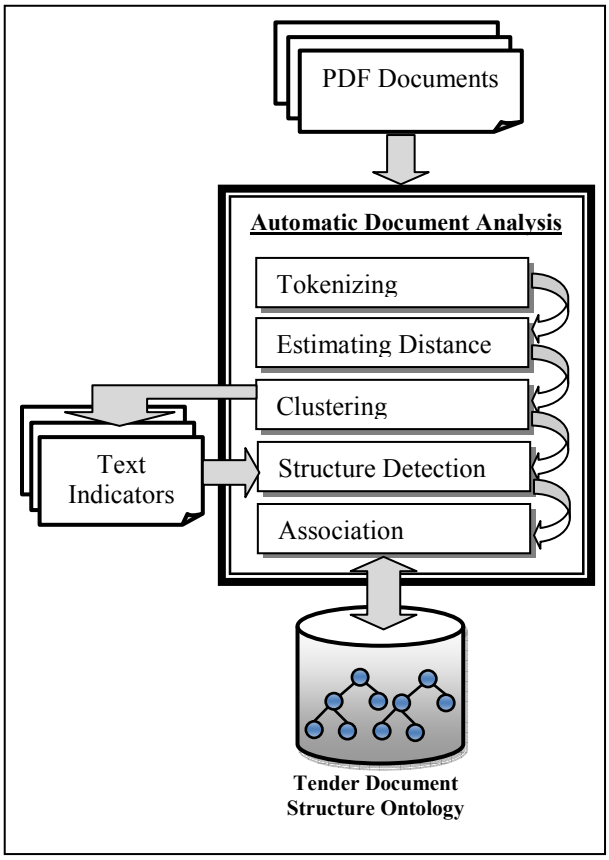


**Figure 1.** Series of Steps in Recognizing Structure and Content of PDF Documents

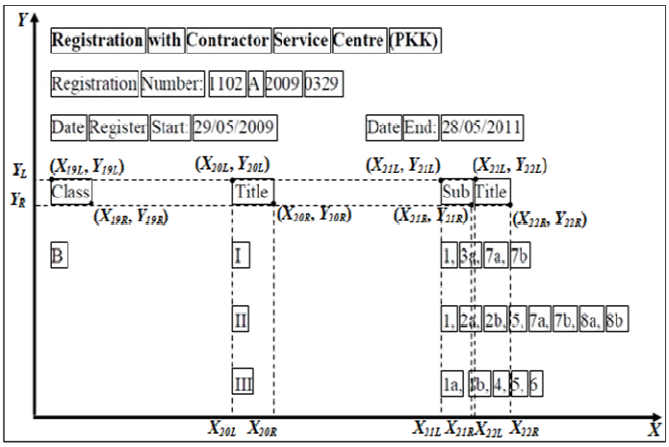


**Figure 2.** Text Elements Defined as Tokens with Upper-Left and Bottom-Right Coordinates

### 3.2 Estimating Distance

Distance estimation is the process to measure space distribution between each token with the closest token to it (nearest neighbour) horizontally and vertically. Space distribution analysis is the most vital stage for computing distance threshold values that will be used in the next step, clustering. Range of distances between tokens are computed separately line by line. This is due to the possibility to have different space distribution between tokens in different horizontal lines when the appearance of tokens on the document are affected by different alignment properties, either left, center, right or justify. Thus, horizontal distance between tokens in the same line are calculated as below.

$$DH_{ij} = \left| X_{iR} - X_{jL} \right|, \text{for } i = 0,1,...,n-1 \text{ and } j = i+1 \quad (1)$$

$DH_{ij}$ represents the distance between token $i$ and token $j$, whilst $X_{iR}$ and $X_{iL}$ denotes the bottom-right X-coordinate of token $i$ and upper-left X-coordinate of token $j$ respectively. The distance estimation is measured by identifying the smallest distance among tokens distributed in the whole page of document. Then, details analysis on the minimum distance is done using statistical analysis. If the difference between minimum distance and other computed distances between tokens that have been compared is significantly produced minimum standard deviation error, then the computed distance is defined as the threshold value. Therefore, the number of threshold values determined is based on the number of horizontal lines identified. Meanwhile, the range of distances between vertical lines are also identified in the same way where the distance

between vertical line and it neighbours are measured as shown in the equation below.

$$DV_{ab} = \left| Y_{aR} - Y_{bL} \right|, \text{for } a = 0,1,...,n-1 \text{ and } b = a+1 \quad (2)$$

$DV_{ab}$ represents the distance between vertical line $a$ and vertical line $b$, whilst $Y_{aR}$ and $Y_{aL}$ denotes the bottom-right Y-coordinate of vertical line $a$ and upper-left Y-coordinate of token $b$ respectively.

### 3.3 Clustering

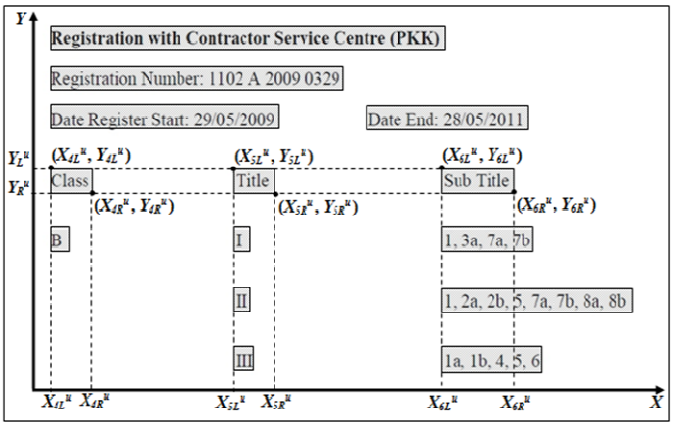


**Figure 3.** Clusters Produced with their Centroid Coordinates

Agglomerative hierarchical clustering algorithm is implemented to group the closest tokens into a cluster in the same horizontal line. The main purpose of clustering is to group the nearest tokens into a similar block. Initially, the process starts with each token as a separate cluster, $c$ and both upper-left and bottom-right coordinates are defined as the cluster centroid. The number of clusters is reduced by merging two clusters that has minimum distance within horizontal threshold value identified in prior step. Then, the updated upper-left ($X_{CL}^{u}$, $Y_{CL}^{u}$) and bottom-right ($X_{CR}^{u}$, $Y_{CR}^{u}$) centroid coordinates are determined for new cluster created as displayed in Fig. 3. The remaining clusters are continuously merged until there is no more non-clustered element left. The clustering process is repeated for other horizontal lines. In other way, similar hierarchical clustering algorithm also applied to cluster the nearest group of tokens vertically. The output from this step is groups of text indicators in which represented by clusters produced.

### 3.4 Structure Detection

Depending on the number of clusters produced in the previous step, PDF document structure is detected for any existing paragraph and tabular layout. At first, each of the horizontal line is tagged either as non-table or table depends on the structure detection rules described as follow. The line is assumed as non-table line when meet one of the following rules; i) if one cluster or one block is created for the line and length of the cluster is greater than half of document length horizontally, or ii) if one cluster is created for line and length of the cluster is smaller than half of document size and is the first line of the document, or iii) if one cluster is created for the line and length of the cluster is smaller than half of document size and the closest above and below lines exist in the same cluster vertically is defined as non-table line, or iv) if one cluster is created for the line and length of the cluster is smaller than half of document size and is the last line of the document and the closest above line exists in the same cluster vertically is defined as non-table line. Otherwise, the line is assigned the opposite tag, table line when meet one of the following rules; i) if more than one cluster are created at the particular line, or ii) if one cluster is created for the line and length of the cluster is smaller than half of document size and the closest

above and below lines exist in the same cluster vertically is defined as table line, or iii) if one cluster is created for the line and length of the cluster is smaller than half of document size and is the last line of the document and the closest above line exists in the same cluster vertically is defined as table line. After performing these rules, paragraph and tabular structure in the document could be identified. Overlapping clusters horizontally and vertically corresponding to the table line identified is considered to reside in the same row and in the same column for the similar table. Row and column headers are detected by matching with the predetermined header keywords. Meanwhile, non-table line identified is considered as paragraph.

### 3.5 Association

The final step is to identify and associate appropriate possible keywords defined in the ontology with the any recognized paragraph and tabular. Initially, text located within the paragraph is defined belong to the paragraph. Text existed within recognized tabular structure is belong to the table itself. Meanwhile, in order to significantly describe about the recognized table, in this case title of table, the nearest paragraph to the tabular structure is considered as possible text that describes about the particular table. Finally all the possible structures recognized are stored in the ontology.

### 3.6 Tender Document Structure Ontology

The ontology is built as the repository to store semantic knowledge on the structure of construction tender document domain by defining information on the document into three different categories such as unstructured (sentences), semi-structured (form-based) and structured (tabular). The purpose of the ontology is to to store construction domain expert knowledge on the structure of tender documents. It also models the keywords as the main sources to extract relevant information.

## 4 EXPERIMENTAL SETUP AND IMPLEMENTATION

The purpose of this experiment is to evaluate the performance of the proposed approach according to precision and recall measurements. A collection of tender documents in PDF for similar building construction project based on Malaysia Construction Tender are used as the experimental data. The total pages of these tender documents is 289 pages. Each page may contains various different layout and structure, for example different size spacing, different type of text alignments (left, center, right, justify). The information on these documents are visually represented in text-based full sentences, form-based and table.

There are two different ways to create PDF document, 1) scanning the existing document using Optical Character Recognition (OCR) and 2) converting documents to PDF using existing tools such as Microsoft Word 2007 onward, PDFWriter, GhostScript and other commercial tools. PDF documents generated from both methods are different in which the former considers the whole content of PDF document as image and the later converts the objects into text and images. In this research, we used PDF documents generated from Microsoft Word 2007. In addition, the experiments are run in Java-based environment and divided into three

strategies according to information types.

## 5 RESULT AND DISCUSSION

Several series of experiments were run to extract the paragraph and tabular structure together with their associated text elements. In order to evaluate the extraction accuracy result, standard information extraction method of precision (*PM*), recall (*RM*) and f-measure have been applied. The standard formula for these measurements are shown below.

$$PM = \frac{R}{R + I} \quad (3)$$

$$RM = \frac{R}{R + N} \quad (4)$$

$$f - Measure = \frac{2 * PM * RM}{PM + RM} \quad (5)$$

*R* is denoted as the number of structure recognized that are relevant, *I* represents the number of irrelevant structure recognized, and *N* is the number of relevant structure not recognized. Table 1 shows the comparison results of these evaluation methods for computerized extraction. Two parameters that have been evaluated are tabular structure (row and column detection) and paragraph detection with their associated text. The evaluation of precision, recall and f-measure have shown significantly good accuracy in detecting relevant information. The precision, recall and f-measure percentage rates for tabular structure detection are 76.8 %, 84.0 % and 80.3 % respectively. Meanwhile, the accuracy of paragraph detection have achieved significantly higher results in which 99.4 % of precision, 96.5 % of recall and 97.9 % of f-measure.

**Table 1.** Comparison Results of Tabular and Paragraph Structure Recognition based on Precision, Recall and f-Measure

| Parameters | Computerized Information Extraction Measurements | | |
|---|---|---|---|
| | Precision | Recall | f-Measure |
| Table | 76.8 % | 84.0 % | 80.3 % |
| Paragraph | 99.4 % | 96.5 % | 97.9 % |

The difference of test accuracy between tabular and paragraph is due to the complex structure of tabular format itself compared to paragraph. The finding shows that proposed approach is significantly capable in recognizing the structure of PDF documents, focusing on tabular and paragraph.

## 6 CONCLUSION

This study proposed an approach for detecting and recognizing PDF document structure, focusing on paragraph and tabular, then associated text using a combination of heuristic, rule-based and predefined indicators. An experimental study involves a collection of construction tender documents. Based on the evaluation measures of precision, recall and f-measure, the result has shown significant performance when recognizing the document structure and associated text. Current work in this paper is based on simple table assumption. Therefore, the future plan of this research is to include complex table (inner columns and rows) analysis.

## 6 REFERENCES

1. ADLIB: White Paper: PDF/Archive – Portable Document Format/Archive. Vol. 2011 (2011)
2. Rosmayati, M., Abdul Razak, H., Zulaiha, A.O., Noor Maizura, M.N.: Ontological-based for Supporting Multi Criteria Decision-Making. In:

Desheng Wen, Zhou, J. (eds.): 2010 2nd IEEE International Conference on Information Management and Engineering, Vol. 1. IEEE Press, Chengdu, China (2010) 214-217
3. Oro, E., Ruffolo, M.: XONTO: An Ontology-Based System for Semantic Information Extraction from PDF Documents. 20th IEEE International Conference on Tools with Artificial Intelligence 2008 (2008) 118-125
4. Klink, S., Dengel, A., Kieninger, T.: Document Structure Analysis Based on Layout and Textual Features. International Workshop on Document Analysis Systems, Rio de Janeiro, Brasil (2000) 41-52
5. Froelich, J., Ananyan, S.: Decision Support via Text Mining. In: Burstein, F., Holsapple, C.W. (eds.): Handbook on Decision Support Systems 1. Springer Berlin Heidelberg (2008) 609-635
6. Liu, Y., Bai, K., Mitra, P., Giles, C.L.: Improving the Table Boundary Detection in PDFs by Fixing the Sequence Error of the Sparse Lines. 2009 10th International Conference on Document Analysis and Recognition, Barcelona, Spain (2009) 1006-1010
7. Jiang, D., Yang, X.: Converting PDF to HTML Approach Based on Text Detection. 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human. ACM New York, NY, USA, Seoul, Korea (2009) 982-985
8. Harvey, G.: Adobe Acrobat 6 PDF For Dummies, Vol. 1. Wiley Publishing, Inc., Indianapolis, Indiana (2003)
9. Zanibbi, R., Blostein, D., Cordy, J.R.: A Survey of Table Recognition: Models, Observations, Transformations, and Inferences. International Journal on Document Analysis and Recognition **7** (2004) 1-16
10. Hassan, T., Baumgartner, R.: Table Recognition and Understanding from PDF Files. International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, Brazil (2007) 1143-1147
11. Yildiz, B., Kaiser, K., Miksch, S.: pdf2table: A Method to Extract Table Information from PDF Files. Indian International Conference on Artificial Intelligence, India (2005) 1773–1785
12. Oro, E., Ruffolo, M.: PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents. 10th International Conference on Document Analysis and Recognition 2009. IEEE Computer Society, Barcelona, Spain (2009) 906-910
13. Schmoekel, I.: PDF-Analyzer Pro 4.0. Vol. 1. Software-Development and Distribution, Achim-Uesen, Germany (2010) 1-11
14. Amyuni, T.: PDF Vol. 2010. Amyuni Technologies Inc., Montreal, Canada (2010)