CONTENT EXTRACTION FROM PDF INVOICES ON BUSINESS DOCUMENT ARCHIVES

R.M.C.V. Bandara

168208N

Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

January 2020

CONTENT EXTRACTION FROM PDF INVOICES ON BUSINESS DOCUMENT ARCHIVES

R.M.C.V. Bandara

168208N

Thesis submitted in partial fulfilment of the requirements for the Degree Master of Science in Computer Science

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

January 2020

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works.

…….. ……..

R.M.C.V. Bandara Date

The above candidate has carried out research for the Master of Science in Computer Science thesis under my supervision.

…….. ……..

Dr. Indika Perera Date

Abstract

Archiving documents is a crucial part on information management, and it will give an organization better control over their information processes. When a business expands, more documents will be produced, and it needs to be carefully handled and tracked to make good use of. Output management systems that are working with ERP systems contains thousands of business documents and Portable document format (PDF) is the common output format for these types of documents. These systems need to execute documents search operations frequently. PDF documents Indexing is a critical part in this context. It will boost document search engine efficiency by cutting search space. Content extraction from PDF documents goes a step further and it will allow more structured search queries.

Extracting the document content from a PDF file is a very important. But this is a very challenging task because PDF is a layout-based format that defines the fonts and locations of the individual character as opposed to the semantic units of the text and their role within the document. In this research I have developed a technique to extract content from a PDF file. We can use it for allow more structured search queries on large document archives in output management systems typically work with world leading ERP systems.

On this research mainly considered on four aspects which are correctly identifying words, word order on a paragraph, clear separation of paragraph boundaries and semantic roles of each word. After extracting content from the PDF file, extracted texts content written to an xml document. XML file contains tags to recognize the pages and rotation angle and number of images on each page. Sample set of PDF invoices extracted and calculated the extracted word percentage to evaluate the accuracy of this technique. This tool hits 94.27% accuracy rate according to the results.

ACKNOWLEDGEMENT

First and foremost, I am deeply grateful for the continuous support, insight, and patience of my supervisor, Dr. Indika Perera: without his invaluable support, this thesis would not have been completsed.

I thank Mr. Ishara Yatawara - software architect at Creative Software, who provided insight and expertise that greatly assisted the research.

Finally, I present my appreciation to my family and my friends who were behind me, encouraging and directing me towards the success of my project.

TABLE OF CONTENTS

DECLARATION i

Abstract ii

ACKNOWLEDGEMENT iii

TABLE OF CONTENTS iv

LIST OF FIGURES vi

LIST OF TABLES viii

1. INTRODUCTION 1

1.1 Overview 1

1.2 Problem 3

1.3 Motivation 4

1.4 Objective 5

2. LITERATURE REVIEW 6

2.1 Automatic indexing of PDF documents with ontologies 6

2.2 Automatic Indexing of Scanned Documents - a Layout-based Approach 8

2.3 Content Extraction of PDF Documents 10

2.4 Diagram Extraction of PDF documents 14

2.5 Benchmark and Evaluation for Text Extraction from PDF 15

3. METHODOLOGY 19

3.1 PDF Components 19

3.1.1 Objects 22

3.1.2 File Structure 29

3.1.3 Document Structure 32

3.1.4 Content Stream 34

3.2 Text in a PDF file 34

3.3 Graphical effect on Text 36
3.4 Text state parameters and operators 38
3.5 Text Objects 42
3.6 Extraction of text content 43
4 IMPLEMENTATION 45
4.1 High level architecture 45
4.2 Getting the last cross reference table (xref) offset 46
4.3 Creating the cross-reference table 47
4.4 Finding the root and creating the page tree 47
4.5 Tokenizing 48
4.6 Token Handling 51
4.7 Object Builder 51
4.8 Reading the content objects 52
4.9 Character Mapping 52
4.10 Character Mapping 54
4.11 Creating Text Rendering Matrix (TRM) 54
4.12 Decomposing Text Rendering Matrix 55
4.13 Reading the virtual coordinate plane 55
4.14 Extracting text to an xml document 55
5 RESULTS AND EVALUATION 56
6 CONCLUSION 65
6.1 Summary 65
6.2 Future Works 68
6.3 Limitations 69
7 REFERENCES 70
APPENDIX 74

LIST OF FIGURES

Figure 1: A sample business document 2
Figure 2: Typical page from a technical manual after AIDAS analysed it.....7
Figure 3: Same template and nearly constant positions of index data.....9
Figure 4: Extraction document through the information extraction system. 10
Figure 5: overview of the processing pipeline 11
Figure 6: System processing steps.....12
Figure 7: The flowchart for text segmentation.....13
Figure *8*: An 8x8 SPAS structure for spatial indexing 15
Figure 9: Output file with 3 paragraphs and ground truth file with 1 paragraph 18
Figure 10: three assignments to evaluation criteria in order to assess O against G ...18
Figure 11: PDF Components.....19
Figure 12: Initial structure of a PDF document.....29
Figure 13: Structure of a PDF document. 32
Figure 14: Glyphs painted in 50% gray 37
Figure 15: glyph outline treated as stroked path 38
Figure 16: Character spacing in horizontal writing.....40
Figure 17: Word spacing in horizontal writing 40
Figure 18: Horizontal scaling.....40
Figure 19: Leading.....41
Figure 20: Text Rising 42
Figure 21: High level architecture.....45
Figure 22: Flow Diagram.....45
Figure 23: Getting the last cross reference table offset 46
Figure 24: Page Hierarchy 47
Figure 25: Tokenizers.....48
Figure 26: Tokenizes the xref objects.....49
Figure 27: Tokenizes the objects.....49
Figure 28: Tokenizes the decoded texts.....50
Figure 29: Tokenizes the Character mappings.....50

Figure 30: Decoded stream 52
Figure 31: Decoded to Unicode stream 53
Figure 32: Coordinate plane.....54
Figure 33: sample input 1.....58
Figure 34: Extracted word percentage.....64

LIST OF TABLES

Table 1: White-Space characters 20
Table 2: Delimiter Characters 21
Table 3: Entries on stream dictionaries 27
Table 4: Entries in the file trailer dictionary. 31
Table 5: Text state parameters 38
Table 6: Text state operators 39
Table 7: Text rendering modes 41
Table 8: Latin-text encoding 44
Table 9: Extracted word percentage 60
Table 10: Words precentage vs Number of files 63