Scientific Journal of Impact Factor(SJIF): 4.14

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Special Issue for ICPCDECT 2016,Volume 3 Issue 1

Information Extraction from Unstructured Document

Swapnali Phadtare[1], Anuja Thube[2], Shubhangi Vahile[3] , Aishwarya Waikar[4]

Department Of Computer, AISSMS's IOIT,

Abstract —*Now a days, PDF (Portable Document Format)is commonly used in industry as a common format for data exchange. Extraction of information from unstructured document gives permission for analyzing and representing in structured format. In this paper we present system for discovering knowledge from PDF and then represent it in EXCEL format .For this conversion first extraction of string contained in PDF is done and then applies different components to express in Excel (the logically structured document).*

Keywords- Information Extraction XY Cut Algorithm Ordering Problem Page segmentation Data mining

I. INTRODUCTION

With the development of information technology and the wide spread use of the internet a large number of electronic documents are stored in Portable Document Format (PDF) is known as a common format of electronic documents. PDF documents are used to store important information relating to products, customer data and corporate knowledge. Meta information such as the document's creator, date of creation or date of modification are further integral parts of a PDF document. PDF documents are often used as "containers" to enable the transfer of text, images, videos and other data to other processes independently of the platforms in use. Information extraction is task of automatically extracting structured information from unstructured and/or semi structured machine readable documents.

Information Extraction means extracting relational elements, relationship between these elements, different attributes that describing elements from unstructured document. Extracting information from unstructured document is challenging task. It requires knowledge about Information Retrieval (IR), Natural Language Processing (NLP), machine learning etc. IR is process of actively seeking out information relevant to a topic of interest.

II.RELATED WORK

The problem of information extraction has received much attention [8]. Our main focus is coverting PDF into Excel . But recent work has started to consider how to extract information from Pdf. Our work fits into this emerging direction, which is described in more detail in[2][3][4]. Extracting high-level document structures from PDF files is a very challenging problem with many potential practical applications. But, such an operation must rely on reliable low-level extraction tools. In our opinion, this problem has been largely underestimated so far.[25] Once we have extracted entity mentions, we can perform additional analysis, such as mention disambiguation Thus, such analyses are higher level and orthogonal to our current work. While we have focused on IE over *unstructured text*, our work is related to the problem of inferring a set of rule to extract information from *PDF[26]* . For extraction may help us develop even more efficient IE algorithms[6].

Using IE algorithm extract information and implement search on it using open source tool like Lucene .Lucene is a free open search information retrieval software library. It is high performance full feature text search engine library, written entirely in JAVA. It contains functionality for document processing, indexing and searching. .Finally, this extracted data convert into Excel format.

III. Method

The main advantage of PDF is portability printing and displaying capability.PDF also contain structured information which is not consider by PDF converters. These converters only consider features like font, color, text etc. We mainly focus on not only this displaying feature but also consider discovering knowledge from PDF.

For converting PDF legacy documents into different forms three steps are performed which are :
1. Information selection and pre-processing

2. Analysis
3. Store this meaningful data in database.

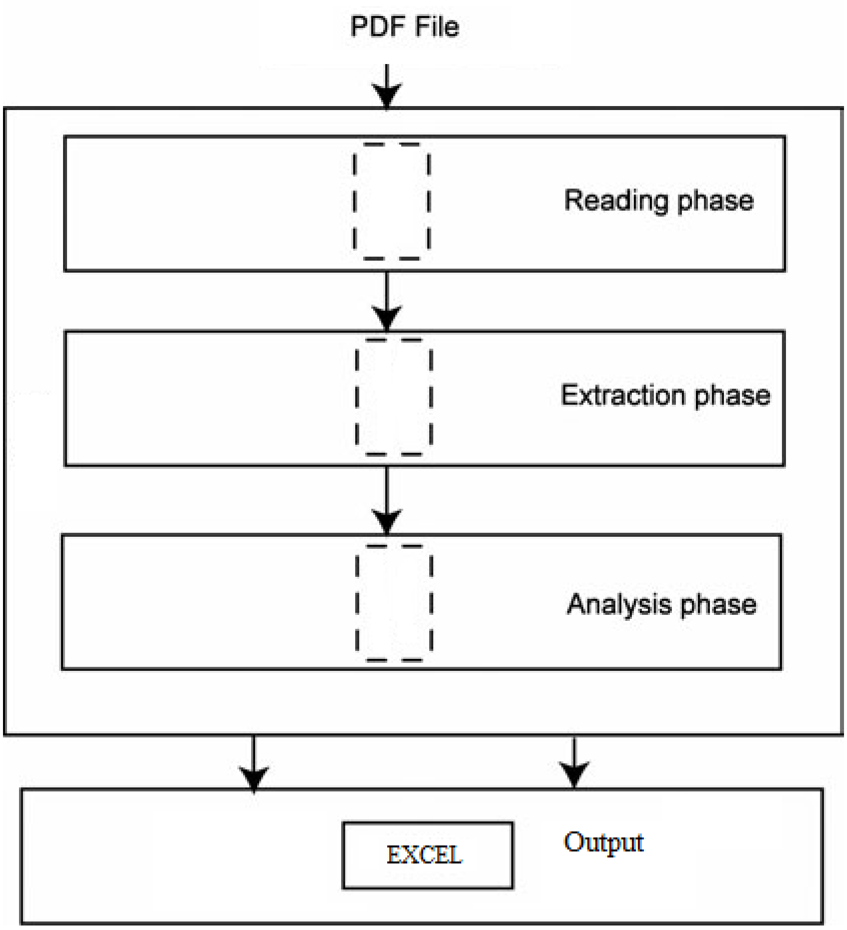

Figure : Block Diagram

A. Information selection and pre-processing

XY-Cut Algorithm

The XY-Cut is a page segmentation method. It is preprocessing method that consists of finding the widest empty rectangle, which crossing the page either vertically or horizontally. The page is then divided in two blocks, which are shrink to fit closely their content. This method segments large regions in document into small regions. This continues till no large-enough block can be found in final result, which become the final segmentation result. From an optimization perspective, the problem of cutting a block can be seen as selecting a series of cuts in order to maximize a score function. The main advantage is to locate the regions of the document where data is present and distinguish is from figures and graphics.

In Some cases choice may not be the optimal one, unless this cut does cross entirely the block, in which case both are equivalent. In more details the block cutting process
1. Given a block to segment, the method enumerates all possible horizontal cuts
2. For each block potentially created by an enumerated horizontal cut, the method enumerates all possible vertical cuts inside it

3. A (possibly empty) set of horizontal cut is chosen so as to make possible the best possible series of enumerated vertical cuts, given the score function.
4. The selected horizontal cuts are performed, and then for each created blocks the set of associated vertical cuts is performed as well.
5. Back to step 1 for each created blocks until no cut is anymore possible

B. Analysis

Implementing search for the extracted data using open source tool like Lucene and grouping this text in a format. Lucene is a free open search information retrieval software library. It is high performance full feature text search engine library, written entirely in JAVA. It contains functionality for document processing, indexing and searching. It is high performance full featured text search engine library written in java. This technology is suitable for nearly any application that requires full text search.

Lucene offers powerful features through simple API such as Scalable, high performance indexing, Powerful, Accurate, Efficient Search Algorithms, Cross-Platform Solution. As Lucene shows wide range of properties so extracted data is analyzed by this tool.

C. Store this meaningful data in database

After analysis, the Lucene would generate keyword terms which would be represented in structured format like EXCEL. For example If PDF contains information about population then after processing table is generated with the fields like state, year, density etc.

IV. Conclusion

Extracting information from PDF files is a very challenging problem as for this we required complete knowledge about natural language processing, information retrieval etc. In this paper we have concentrated on XY-cut algorithm. We have described a ordering method that grounds on the XY-cut method. Ordering method separates lines to help the human reader and those lines also conduct to the

appropriate cutting and hence order. This allows us to split a block into sub-blocks in an optimal manner, while encapsulating the guiding heuristic in a score function. Under the same approach, this function may eventually include additional relevant features that is analysis of data and then evaluating it. This evaluation gives structured data.

.

REFERENCES

[1] Raymond J. Mooney and Razvan Bunescu," Mining Knowledge from Text Using Information Extraction", SIGKDD Explorations.
[2] Jöran Beel, Bela Gipp, Ammar Shaker and Nick Friedrich , "SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size)", *Proceedings of the 14th European Conference on Digital Libraries (ECDL"10),* volume 6273 of Lecture Notes of Computer Science (LNCS), September 2010. pp- 413–416, Glasgow (UK), Springer.
[3] Qingzhao Tan, Prasenjit Mitra,C. Lee Giles, "Metadata Extraction and Indexing for Map Search in Web Documents", *Proceeding of the 17th ACM conference on Information and knowledge management,* ACM New York, NY, USA ©2008,pp- 1367-1368
[4] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 2013
[5] Jean-Luc Meunier " Optimized XY-Cut for Determining a Page Reading Order", Proceedings Eighth International Conference on Document Analysis and Recognition, 2005.
[6] JL Meunier, Optimized XY-Cut for Determining a Page Reading Order, ICDAR 2005
[7] X. Lin, "Text-mining Based Journal Splitting", Proceedings of the Seventh International Conference on Document Analysis and recognition, ICDAR'03, 2003.
[8] Fei Chen , AnHai Doan , Jun Yang, Raghu Ramakrishnan, "Efficient Information Extraction over Evolving Text Data"

[9] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, and Shivakumar Vaithyanathan, " An Algebraic Approach to Rule-Based Information Extraction",

[10] Jaekyu Ha, R.M. Haralick, I.T. Phillips, “Recursive X-Y cut using bounding boxes of connected components, International Conference on Document Analysis and Recognition”, ICDAR 1995

[11] Yasuto Ishitani, “Document Transformation System from Papers to XML Data Based on Pivot XML Document Method”, International conference on document analysis and recognition, ICDAR 2003

[12] A. K. Jain, M. N. Myrthy, and P. J. Flynn. “Data clustering: A survey”. ACM Computing Survey, 31(3):264--323, 1999.

[13] R. Cattoni, T. Coianiz, S. Messelodi, C.M. Modena: “Geometric Layout Analysis Techniques for Document Image Understanding: a Review”, ITC-IRST Technical Report #9703-09

[14] A. Antanacopoulos, B. Gatos and D. Karatzas, “ICDAR 2003 Page Segmentation Competition”, *ICDAR2003*, Edinburgh (Scotland), August 2003, pp. 688-692.

[15] BCL, http://www.bcltechnologies.com/document/index.asp

[16] R. Cattoni, T. Coianiz, S. Messelodi and C.M. Modena, “Geometric layout analysis techniques for document image understanding a review”, *Technical report, IRST*, Trento, Italy, 1998.

[17] Glance, http://www.pdf-tools.com/en/home.asp

[18] K. Hadjar, O. Hitz and R. Ingold, “Newspaper Page Decomposition using a Split and Merge Approach”, *ICDAR’01*, Seattle (USA), September 2001, pp. 1186-1189.

[19] K. Hadjar and R. Ingold, “Arabic Newspaper Page Segmentation”, *ICDAR’03*, Edinburgh (Scotland), August 2003, pp. 895-899.

[20] K. Hadjar, O. Hitz, L. Robadey and R. Ingold, “Configuration REecognition Model for Complex Reverse Engineering Methods: 2(CREM)”, *DAS’02*, Princeton, NJ (USA), August 2002, pp. 469-479.

[21] R.M. Haralick, “Document image understanding: Geometric and logical layout”, *Proc. Internet. Conf. On Computer Vision and Pattern Recognition*, 1994, pp. 385-390.

[22] O. Hitz, L. Robadey and R. Ingold, “An architecture for editing documents recognition results using xml technology”, *DAS’2000*, Rio de Janeiro (Brazil), December 2000, pp. 385- 396.

[23] J. Hu, R. Kashi, D. Lopresti, G. Nagy and G. Wilfong, “Why table ground truthing is hard”, *ICDAR’01*, Seattle (USA), September 2001, pp. 129-133.

[24] Jonathan H. Aseltine ”WAVE:An Incremental Algorithm for Information Extraction” AAAI Technical Report,1999

[25] Karim Hadjar, Maurizio Rigamonti, Denis Lalanne and Rolf Ingold **“**Xed: a new tool for eXtracting hidden structures from Electronic Documents”

[26] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, and Shivakumar Vaithyanathan “An Algebraic Approach to Rule-Based Information Extraction”