INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 2)

Available online at: www.ijariit.com

Result extraction from searchable PDF

Manish Yadav
manish.manish.yadav52@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Harshit Virkar
hvirkar36@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Ishan Tipnis
ishantipnis23@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Rohan Gaikwad
rohangaikwad6496@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Namita Pulgam
ashwinigudewar@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

Kamlesh Nenwani
nenwani.kamlesh@gmail.com
Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra

ABSTRACT

Digital documents present knowledge in most areas of study, exchanging and communicating information in a portable way. These digital repositories help in providing efficient retrieval of the stored data thus making it an important tool. Systematic extraction of data from digital document helps in mitigating the tedious work of manual data entry and facilitates effective analysis of generated structured data. The proposed system aims to extract a student's final result from the digital copy of the result sheet (PDF file) and then storing it in a centralized database which reduces the tiresome manual work needed to store records of each and every student. It provides a novel method of detecting text which is organized in tabular format and retaining the structural information after text recognition with good accuracy.

Keywords: *XML, Searchable PDF.*

1. INTRODUCTION

Portable Document Format (PDF) since its inception in 1993 has become the most widespread document exchange platform due to high level of interoperability between texts and graphics. A major reason for its universal acceptance is the fact that the document is rendered identically to both sender and receiver. It is this tradeoff between the comfortable reading it provides and a painful manipulation of data it requires which makes data extraction from PDF both essential and tedious. Though tables are the most prevalent means of communicating structured data, lack of uniformity in its presentation makes table extraction extremely difficult especially in PDF documents. The main issue in detecting tables is that they are mixed with other elements at the same time lack tags which can help in accurate identification. This is due to, (1) Presence of open tables; (2) Merged rows or columns; (3) Table content laid out across several pages; (4) Missing values which makes predicting layout difficult.

Document processing involves the conversion of typed and handwritten text on paper-based and electronic documents (e.g., scanned image of a document) into electronic information utilizing one of, or a combination of, intelligent character recognition (ICR), optical character recognition (OCR) and experienced data entry clerks.

There are three types of PDF files namely (1)True PDF File which is originally generated from a computer by Word, Excel, InDesign, Illustrator and are built of code that allows them to be viewed and read exactly as they were originally created. (2)The Scanned PDF File contains no electronic code to maintain its integrity and is no more than an image. (3)Searchable PDFs usually result through the application of OCR (Optical Character Recognition) to scanned PDFs or other image-based documents. During the text recognition process, characters and the document structure are analyzed and read. A text layer is added to the image layer, usually placed underneath.

True PDFs give highly accurate results as they can be easily converted into XML due to the presence of both characters in the text and the meta-information having an electronic character designation. Contrary to this Scanned PDFs need to be converted into Searchable format for proper XML parsing. This can be done through various image processing operations for optimal detection and recognition of tables.

As the whole process to view ones result is complex as well as quite tiresome for the college staff to enter the data manually and verifying the same. We propose a solution to automate this method which will not only ease the process of viewing the results for the student but also substantially reduce the work of the college staff. The admin of the college just needs to upload the searchable pdf document obtained by the college from the University to our web application. The whole pdf file will be converted into XML format and with the help of this XML document, the data will be extracted from the text tag and stored in the NoSQL database. Thus saving a lot of time and effort of all the stakeholders involved. This acquired database can be used for further result analysis like to know the passing percentage of the college, or to find the highest marks scored by any student in a particular subject and so on. The rest of the paper is organized as follows. Section reviews several relevant studies in table detection and extraction, especially those on PDF files. Section III firstly gives an overview of the proposed solution and then presents each step in detail. Experimental results are demonstrated in Section IV. Finally Conclusion and future work is mentioned in Section V.

2. LITERATURE REVIEW

This section includes the current knowledge along with substantive findings, as well as theoretical and methodological contributions related to our topic. It is basically an evaluative report of information found in the literature relevant to ta-ble detection, text recognition and various image processing operations to complement result extraction process, thereby increasing its efficiency and accuracy, as well as maintaining the structural integrity of the generated output.

R. Zanibbi et al.[1] characterize the various table recognition methods as combination of observations, transformations and inferences. Where observations collect and compute data of physical structure, logical structure or already present de-scriptive statistics used in decision making in a table recognizer. Transformations restructure prevailing observations to emphasize features of a data set, to making subsequent observations easier or more reliable. These may be logical (tree and graph), physical (Hough) or simple preprocessing operations like binarization, resampling, compression and mathematical morphology. Inferences (Classifiers, Segmenters, and Parsers) decide whether or how a table model can be fit a document, through the generation and testing of hypotheses.

Wang, Y. [2] has categorized the table processing approaches into Predefined Layout-based approach where several templates of possible table structures are created and the input documents or portions of it which fit a template are identified as tables; Heuristic-based approach which uses a set of rules for decision making; and Statistic or optimization-based approach which produces statistical measures by offline training and the consequent parameters are used for decision making.

Ermelinda Oro, et al.[3] propose a heuristic approach for table recognition and extraction from PDF documents named PDF-TREX which plots tables present in the PDF documents as a 2D grid on a Cartesian plane and extracts them as a group of individual cells by utilising the aforementioned 2D coordinates. An agglomerative hierarchical clustering algorithm is used to build segments and blocks that form table structure. The output is generated as an XML representation of extracted cells which may be used for further processing.

Martha O. Perez-Arriaga, et al.[4] has commented on the process of Table detection and recognition through systematic identification and extraction of the cells contained in a table using Table Organization (TAO), a processing tool based on the k-nearest neighbor method and layout heuristics. Here a PDF document acts as the input for TAO, which is subsequently converted into an XML format using PDFMiner which generates separate XML tags for each component of the document. While the detection module identifies table candidates through structural analysis, the extraction module uses the information from probable table candidates to find individual cells and their content by comparing the text lines. The tables are also augmented with information related to metadata and position within the document finally producing a JSON document as output. This top-down approach helps in identifying structure before analyzing the specific lines that may belong to a table as compared to the line-by-line search of PDF-TREX.

Yildiz et al. [5] describe PDF2Table the first open source tool to specifically extract tables from PDF. It is based on an heuristic approach that performs two main tasks: table recognition, in which information organized in tabular structure are recognized, and table decomposition in which recognized elements are assigned to a table model. It uses pdftohtml tool to return text chunks and their absolute coordinates in the PDF file in the same order as they were inserted into the original file. The output of this preprocessing step is fed into the algorithm which works on the principle that table will have more than one column and finally produces an XML file as an output. But it is plagued with many shortcomings biggest being the over reliance on pdftohtml tool. Thus if this tool produces incomplete or incorrect information the rest of the steps become meaningless. It is also unable to distinguish between hidden tables (tables which are not labeled as such in the original file) and real tables.

S. Deivalakshmi et al. [6] showcased a fast, language independent skilled technique for table structure detection and its content extraction from a scanned document image based on morphological operation, connected components and labeling. A major issue in this method is that it is not applicable in case there are no lines on the input document or the documents with tables are made up of only one kind of lines.

Hong Tai Tran et al. [7] presents a novel method of extracting table cells from document images. While, ruling lines are used in standard cases, in their absence the input image is divided into text and lines images. And then, the table is decomposed from those two images separately.

The available literature on this topic helped us understand that though table has a typical and seemingly simple format, its detection and processing from a PDF is quite difficult and requires computationally complex methods. Also processing scanned or image only PDF is a tedious job due to the uncertainty in the functioning of OCR tools. Contrary to this using XML format helps in preserving the structural information and also provides good accuracy.

3. PROBLEM DEFINITION

The proposed Result Extraction tool is a web application that provides for efficient detection and extraction of the result information present in tabular format from the PDF and storing it in a database in a logical and structured form. The aforementioned database shall act as a repository of all result related data and help in creating a reliable, intuitive and user friendly result analysis tool.

4. METHODOLOGY

In this section, the complete flow of the proposed system is explained. The proposed approach aims to extract tabular data from a digital document and store it under the appropriate fields in a database. This method consists of following steps: In the first step we need to upload the searchable pdf document into the web application which will convert it into XML file, in the next step elimination of redundant static text regions takes place, following it, we logically group, segregate and extract data from the XML. Succeeding the above step, addition of acquired extracted data into the database as well as generation of an Excel sheet takes place which will result in easier analysis of the data.

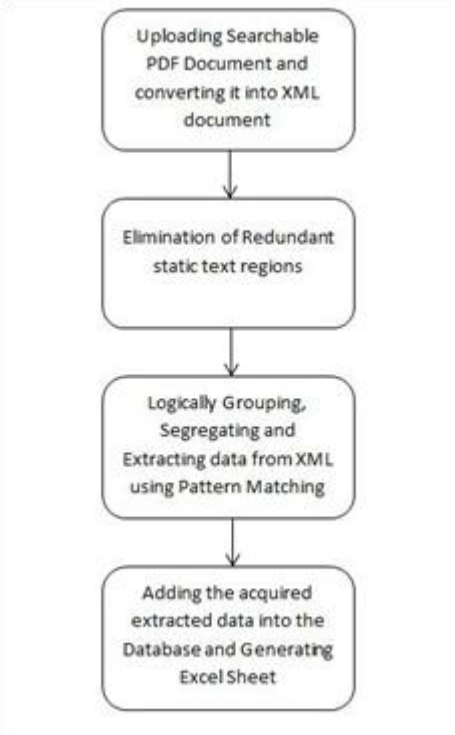


Fig. 1: Flow of the system

1- Upload your searchable pdf document and its conversion to XML file our proposed work will be a web application which will have an upload button in it. Only the admin of the college will have the rights to access it. After verification of the admins credentials, the PDF will be uploaded. After uploading the PDF document, it will be converted into an XML file which can be used for further analysis.

2- Elimination of redundant static text regions in this step, all the needless and undesirable text regions are removed from the file. Primarily, all the text tags which contain white spaces or the tags which are empty are removed programmatically. This process is done to keep up with the time complexity of the overall process and maximize the accuracy of the overall output.

3- Logically grouping, segregating and extracting data from XML The text is segmented logically with the help of Regular

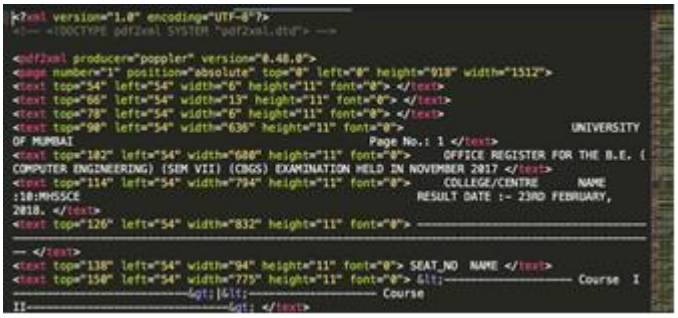


Fig. 2: XML document of Result

Fig. 3: Redundant data in the XML document

Expression. A Regular Expression also known as Rational Expression is a sequence of characters that define a search pattern which is usually used by String Searching Algorithms for find or find and replace operations on strings. For logically segmenting the data, separators like white spaces and — are identified. Logically similar data are grouped to form an object. These objects are the segregated data which forms a cluster of vivid as well as varied objects.

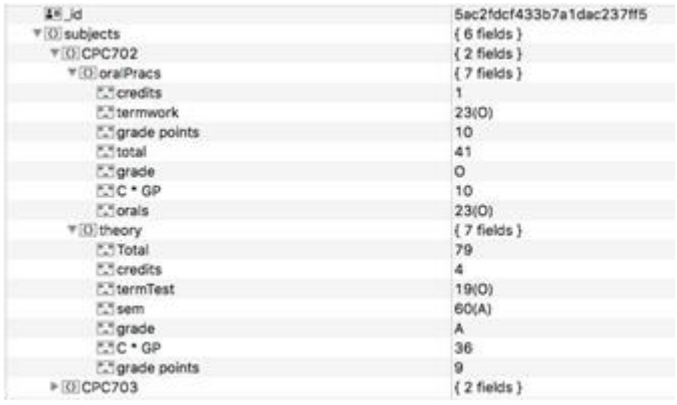
4- Adding the acquired extracted data into the database and Generation of an Excel Sheet the objects which are obtained in the previous step are added to the NoSQL database. Each object forms the student record. This database has a separate field for theory, practical, termwork, and orals mark obtained by the student making it easier for them to view their results. In addition to this, an Excel sheet is also generated which can be used for further analysis of result like knowing the highest percentage scored by an individual, the overall passing percent, number of students failed, etc.

5. EXPERIMENTAL RESULTS

The proposed system extracts the results found in a tabular format and stores it in the database in a structured and logical way. The data is segregated in an organized manner by creating records of each student and mapping their seat numbers. Both combined and individual results of each subject are produced which help in further result analysis. Figure (4) shows the database and created records.

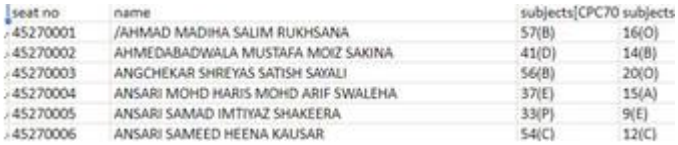
Figure(5) depicts the database converted into Excel format for better understanding and easier result analysis through segregation of data based on subjects as well as score.

This data also acts as a base for detailed result analysis. An instance of which is shown in figure (6) where the total number of passed and failed students has been found out. Also the highest grade point has been retrieved.



_id	5ac2fdcf433b7a1dac237ff5
subjects	{ 6 fields }
CPC702	{ 2 fields }
oralPracs	{ 7 fields }
credits	1
termwork	23(O)
grade points	10
total	41
grade	O
C * GP	10
orals	23(O)
theory	{ 7 fields }
Total	79
credits	4
termTest	19(O)
sem	60(A)
grade	A
C * GP	36
grade points	9
CPC703	{ 2 fields }

Fig. 4: Result Database



seat no	name	subjects[CPC70	subjects
45270001	/AHMAD MADIHA SALIM RUKHSANA	57(B)	16(O)
45270002	AHMEDABADWALA MUSTAFA MOIZ SAKINA	41(D)	14(B)
45270003	ANGCHEKAR SHREYAS SATISH SAYALI	56(B)	20(O)
45270004	ANSARI MOHD HARIS MOHD ARIF SWALEHA	37(E)	15(A)
45270005	ANSARI SAMAD IMTIYAZ SHAKEERA	33(P)	9(E)
45270006	ANSARI SAMEED HEENA KAUSAR	54(C)	12(C)

Fig. 5: Result in Tabular format

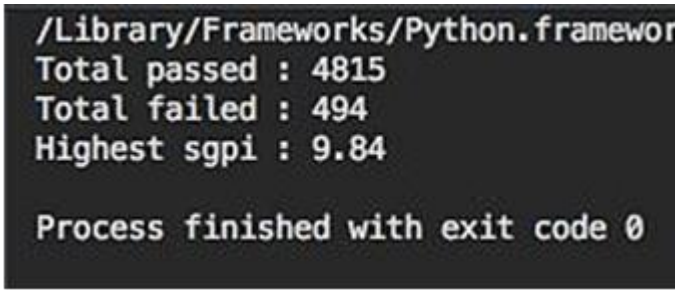


Fig. 6: Result Analysis

6. CONCLUSION

The proposed system aims to automate the process of manual data entry for each student thereby easing the tiresome manual work as well as provides students a simple way to view and understand their results. If the pdf file is in scanned or non-searchable format, OCR (Optical Character Recognition) is used to extract the data from the pdf. Even the modern OCR Engines provides a comparatively low accuracy when it comes to table extraction and at the same time the structural information of tabular data is lost after text recognition. In such a case, detecting and tagging text regions in tabular data helps in preserving structural information of the tabular data after text recognition and facilitates storing of recognized text under an appropriate field in a database. Whereas in case of a true or searchable pdf, it can be converted into an XML document and by traversing through it text is extracted from the text tags. The generated text is then segmented logically and grouped to form an object. Each object forms a student record which is stored in the database sequentially. The proposed system even facilitates result analysis on the acquired database, like knowing the passing percentage, highest SGPI scored, highest marks scored by an individual in any subjects, number of students absent can be acquired through this database easily and accurately. Thus processing of searchable pdf is much more efficient and gives higher accuracy.

7. REFERENCES

[1] R. Zanibbi, D. Blostein and J.R. Cordy,"A Survey of Table Recognition: Models, Observations, Transformations, and Inferences", Document Analysis and Recognition March 2004, Volume 7, Issue 1
[2] Yalin Wang, M. Haralick, et al., Document Analysis: Table Structure Understanding and Zone Content Classification, PhD thesis, Washington University (2002)
[3] Oro, Ermelinda and Massimo Ruffolo. , PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents. 10th International Conference on Document Analysis and Recognition (2009)
[4] Perez-Arriaga, Martha O. et al. TAO: System for Table Detection and Extraction from PDF Documents. FLAIRS Conference (2016).
[5] Yildiz, Burcu et al. pdf2table: A Method to Extract Table Information from PDF Files. IICAI (2005).
[6] S.Deivalakshmi, K.Chaitanya and P.Palanisamy,"Detection of Table Structure and Content Extraction from Scanned Documents", Interna-tional Conference on Communication and Signal Processing, April 2014.
[7] Hong Tai Tran, Tuan Anh Tran, In Seop Na, Soo Hyung Kim , "Cell Decomposition for the Table in Document Image Based on Analysis of Texts and Lines Distribution", Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference, August 2016.
[8] B. Gatos, D. Danatsas, I. Pratikakis and S. J. Perantonis,Automatic Table Detection in Document Images, International Conference on Pattern Recognition and Image Analysis., ICAPR 2005: Pattern Recognition and Data Mining
[9] Naganjaneyulu, G. V. S. S. K. R. et al. A multi clue heuristic based algo-rithm for table detection. 2016 IEEE Region 10 Conference (TENCON) (2016): 1246-1249.
[10] Corrła, Andreiwid Sheffer and Pr-Ola Zander. Unleashing Tabular Con-tent to Open Data: A Survey on PDF Table Extraction Methods and Tools. DG.O (2017).

