Extract and Analyze Data from PDF File and Web : A Review

Darshana Jadhav [1], Dhanashree Jadhav [2], Pooja More [3], Harshali Nikam

[1] Darshana Jadhav , Dept. of computer Engineering, MET, Nashik

[2] Dhanashree Jadhav , Dept. of computer Engineering, MET, Nashik

[3] Pooja More, Dept. of computer Engineering, MET, Nashik

[4]Harshali Nikam, Dept. of computer Engineering, MET, Nashik

Assistant Professor : Ms.Tusharsaheb Patil

Abstract - *Current survey done on today's scenario shows, result gadget declared by Universities(eg. Pune Uni.) for engineering is in PDF file format. The PDF data contents detail such as seat no, centre, permanent registration no.(PRN), Name, Subjects, Marks, etc. Presently PDF file is extracted in excel file format, this conversion is done in order to extract various reporting formats required by department/college/university at various level. Thus, it involves somewhat manual process. However, all these operation have certain limitations such as semi-automated process, no GUI present, SMS gateway is not support, E-mail gateway is not supported, and mainly graphical analysis of data is not available. On the basis of survey done, we came across existing applications which are semi-automated or automated with some restrictions which does not allow full automation of result analysis in proper format. Thus none of the applications supported the full automation. To overcome above said drawbacks, we proposed a new system for result analysis, which is automated with features like Auto-output generation in different database format like excel, PDF, Mysql for further compatibility with other ERP system as per user selection, active SMS gateway, active Email gateway, interactive and user friendly GUI, graphical result analysis with text. In Proposed system we have targeted the limitations to provide effective solution for result analysis. This system will also work on current grade system. Where we are going to maintain database of students which will show whole status of students. Automated solutions provided by the system will make exam department activities more efficient by covering most of the important drawbacks of manual system, namely speed, precision and simplicity. It will also work as a generalized system to support any type and format of PDF file. A centralized system will ensure that the activities in the context of an examination can be* managed effectively, *while also making it more accessible and convenient for both staff and students.*

***Key Words*: Information Extraction, Pattern Matching, Data Mining, Web Mining.**

1.INTRODUCTION

Result evaluation and analysis requires plenty of manual work. so in order to reduce this issue we need system which will support automation. Our system will work for university results. Nowadays in most of the engineering colleges , the traditional method carried out by the colleges is to fill the data within excel sheet manually for each student from the pdf file provided by the university. There are so many formulas for categories the things like toppers, pass, fail, droppers, etc. This is a complete manual process where chances of mistakes are so high. Similarly in diploma colleges results are declared online, so data is taken from web and fill into excel sheet manually and accordingly the data evaluated and analyzed as per requirements of result reports. This process is actually a very time consuming. Thus in order to fill ease the people doing this analysis, we have propose one system which would automate the process of result evaluation and analysis. This system take the input as pdf file provided by university and save into database, once the data get store into database we can use the data to get the information using various queries.

2. LITERATURE SURVEY

In Existing System the data sort and analyze by manual processes. User has to copy/paste the pdf file into excel sheets and have to manually sort it to rank students. Proposed system will be used to automate these processes. Several researchers work on the topic of extracting require data from unstructured data such as PDF. Here we are going describe the tools which are closely related to proposed system in this section. In reference [1] the authors used the PDF-Box technique to extract references from PDF which converts the PDF data into text and get the require information from data. In reference [2] author used LA-PDFText technique which is a command line utility to extract text from PDF just by providing path of PDF file. In [3] author uses a technique for extraction of data from the structured web pages. In reference [4] author uses a technique called tag injection which inserts format information into text document which is in the form of tags. It helps to transform a text into semi structure data, their is complete details are discussed about data extraction .

3. PROPOSED SYSTEM

Following figure shows the detailed view of the proposed system :

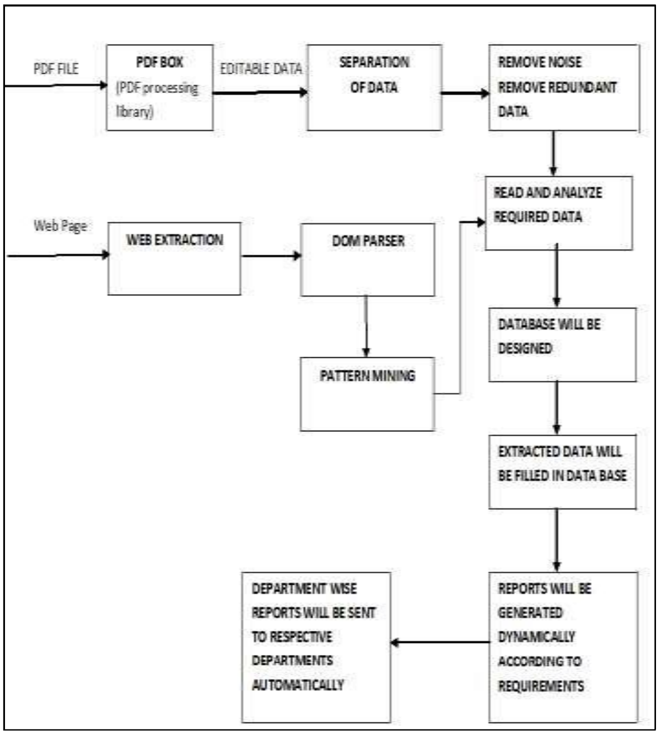
PDF FILE
PDF BOX
(PDF processing library)
EDITABLE DATA
SEPARATION OF DATA
REMOVE NOISE
REMOVE REDUNDANT DATA
READ AND ANALYZE REQUIRED DATA
Web Page
WEB EXTRACTION
DOM PARSER
PATTERN MINING
DATABASE WILL BE DESIGNED
EXTRACTED DATA WILL BE FILLED IN DATA BASE
REPORTS WILL BE GENERATED DYNAMICALLY ACCORDING TO REQUIREMENTS
DEPARTMENT WISE REPORTS WILL BE SENT TO RESPECTIVE DEPARTMENTS AUTOMATICALLY

Fig: Detailed View

3.1 PDF Box :

PDF file is input for the system, so system has to first extract data from PDF files. Here the PDF file is result gadget provided by the Universities. so it does not contain any diagram or images. To extract data from PDF files, we are going to use PDF box technique.PDF box is PDF processing library, it supports development and conversion of PDF documents in addition it also provides command line utility for performing various operations actually. PDF box has ability to quickly and accurately extract the contents from PDF documents. To use PDF box technique, we have to include iTextSharp package. iText provides API in languages such as .net, android, JAE, java developers to provide enhancement to their application with a PDF functionality. It provide functionalities such as PDF generation, PDF manipulation, and PDF form filling. After including the package, *PdfReader* is used to read the PDF file and then *PdfTextExtractor* is used to extract the portable document data.

3.2 Sorting of data :

Text extracted from PDF files is stored in text file. Proposed system categories the data according to each department. This separation is done by string manipulation operations.

3.3 Remove Noisy/Redundant Data :

After getting the essential data from the extracted PDF data, and filtering the data which is not required. For this purpose, we are using parsing technique which will help us to do parsing line by line. Also the PDF contain lots of redundant data E.g. Input PDF file contain same subject list for each student for his/her of particular department. Then such redundant data is also removed and only single copy of data is stored in the database system.

3.4 WEB Extraction :

WEB extractor recognize the relevant data from the web page and extract two different types of data out from it one is source code and another is plain text displayed on web page.

3.5. DOM Parser for Web Mining:

DOM is Document Object Model usually used for organize the nodes into tree structure extracted from web pages.

3.6 Pattern Mining :

System uses pattern mining method to find the essential data from extracted document. The extracted plain text by the web extractor is checked this the specified pattern and mined the data accordingly.

3.7 Read and Analyze required data :

After elimination the noisy and redundant data, system has need actual data . Then this data is accessed for each student. Analysis of each student data is to be done by the system. For the first time system will divides the department then reading the subject list of each department, seperating subjects into theory, practical, term-work and oral wise, online exam and insem exam and to generate the final result of every individual . Also system read personal information of each student from text extracted from PDF.

3.8. Database designed and extracted data filled in the system :

All gathered data which is useful need to be store into the database system. Thus system designs database dynamically by reading the contents from pdf file. After database is designed, department wise tables are generated. Then in tables analyzed data will be store.

3.9. Reports generation:

Reports are generated using the data is stored in the database. The result reports will be generate by means of requirements. The reports like college topper, department wise topper, subject wise topper, ATKT's, dropper student, etc. System will generate result reports which are send via mail to respective department/students.

4. CONCLUSIONS

System will sort all the data according to students marks and grades if requested by user, for this we use data mining techniques ,PDF extraction, data fetching and sorting techniques, which will make user to simplify the data easily and make result reports accordingly along with graphical representation(using pie charts and graphs). It will become convenient for students to receive results through SMS and Email gateways. By this way result data will be organized well , which becomes easy to manage the result records.

ACKNOWLEDGEMENT

We express our sincere gratitude to Prof. Mr. Tusharsaheb Patil (Assistant Professor, MET BKC IOE) for his support and guidance. We would also like to thank Prof. Mr. Pankaj Deore (Asst. Professor, MET BKC IOE) for his valuable words of advice. We are also extremely grateful to our respected H.O.D. Dr. M. U. Kharat and Principal Dr. V. P. Wani for providing all facilities and every help for smooth progress of project work. We are thankful for our family members and friends for motivating us.

REFERENCES

[1]A Strategy for Automatically Extracting References from PDF Documents*.* *Neide Ferreira Alves,* Universidade do Estado do AmazonasManaus, Brazil *Rafael Dueire Lins*, Universidade Federal de Pernambuco Recife, Brazil *Maria Lencastre,* Universidade de PernambucoRecife.

[2] Automatic classification of scientific papers in PDF for populating ontologies. Juan C. Redon-Miranda, Julia Y. Arana-Llanes, Juan G. González-Serna and Nimrod González- Franco Department of Computer Science National Center for Research and Technological Development, CENIDET Cuernavaca, México {juancarlos, juliaarana, gabriel,

[3] HWPDE: Novel Approach for Data Extraction from Structured Web Pages .Manpreet Singh Sehgal Department of information Technology, Apeejay College of Engineering, Sohna, Gurgaon Anuradha PhD, Department of Computer Engineering, YMCA University of Sc. & Technology, Faridabad

[4] A new method of information extraction from pdf filesFANG YUAN1,2, BO LIU College of Mathematics and Computer Science, Hebei University, Baoding, 071002 P.R.China College of Information Science and Engineering, Northeastern University, She0nyang, 110004 P.R.China.

BIOGRAPHIES

Darshana Jadhav Pursuing her computer degree course in MET's Institute of Engineering, Nashik. Her interest include database system.

Dhanashree Jadhav Pursuing her computer degree course in MET's Institute of Engineering, Nashik. Her interest include database system.

Pooja More Pursuing her computer degree course in MET's Institute of Engineering, Nashik. Her interest include database system and data mining, web mining .

Harshali Nikam Pursuing her computer degree course in MET's Institute of Engineering, Nashik. Her interest include database system and web mining.