www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 4 April 2015, Page No. 11193-11196

Extracting and Mining Of Data From PDF and WEB

Pooja Tajane, Pranjal Gadakh, Nisha Shelar, Madhuri Javare

Department of Computer Engineering, K.K.Wagh Institute of Engineering & Education Research, Nasik, Maharashtra
Email: tajane.pooja@gmail.com
mepranju@gmail.com
nishushelar@gmail.com
madhurijavare@yahoo.com

ABSTRACT

In most of the Universities, results are published on web or send via PDF files. Currently many of the colleges use manual process to analyze the results. Sadly the college staff has to manually fill the student result details and then analyze the rankings accordingly. Our proposed system will extract the data automatically from PDF and web, create dynamic database and analyze data, for this system make use of PDF Extractor, Pattern matching techniques, data mining, Web mining technique and sorting technique.

Keywords- Information Extraction, Pattern Matching, Data Mining, Web Mining.

I. INTRODUCTION

Result analysis requires large amount of manual work. Our system works for Pune university Engineering colleges And Mumbai University Diploma Colleges results. In most of the engineering colleges, the traditional method carried out by the colleges is to manually fill the data in excel sheet of each student from the gadget provided by the university and then using some formulas for various analysis like toppers, droppers, ATKT etc. This method consumes plenty of time and chances of human mistake are very high. Similarly In diploma colleges also manually data from web is filled into the excel sheets and accordingly results are analyzed. Thus in order to relax the people doing this analysis, we have proposed a system which would automate the process of result analysis. This system take input as pdf by Pune university (Gadget) and web pages by Mumbai university, automatically stores the data into the database ,once the database is created we can extract various information from that data using various queries .

II. LITERATURE SURVEY

In Existing System manual sorting and analyzing of data is to be done. User has to read PDF file and have to manually rank students and also students have to go to website and search their score. To automate the above process, proposed system is used.

Several researchers worked on the topic of extracting require data from unstructured data such as PDF. This section described the tools which are closely related to proposed system.

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 data. In reference [2] author used LA-PDFText technique which provides a command line interface to extract text from PDF just by providing path of PDF file. In [3] reference author describes the technique which extracts the web page data from hidden 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. In reference [5] author discussed about technique which is used to extract the figures in Portable Document using PDF box or other PDF processing libraries.

III. PROPOSED SYSTEM

Basic block diagram of proposed system as follow:

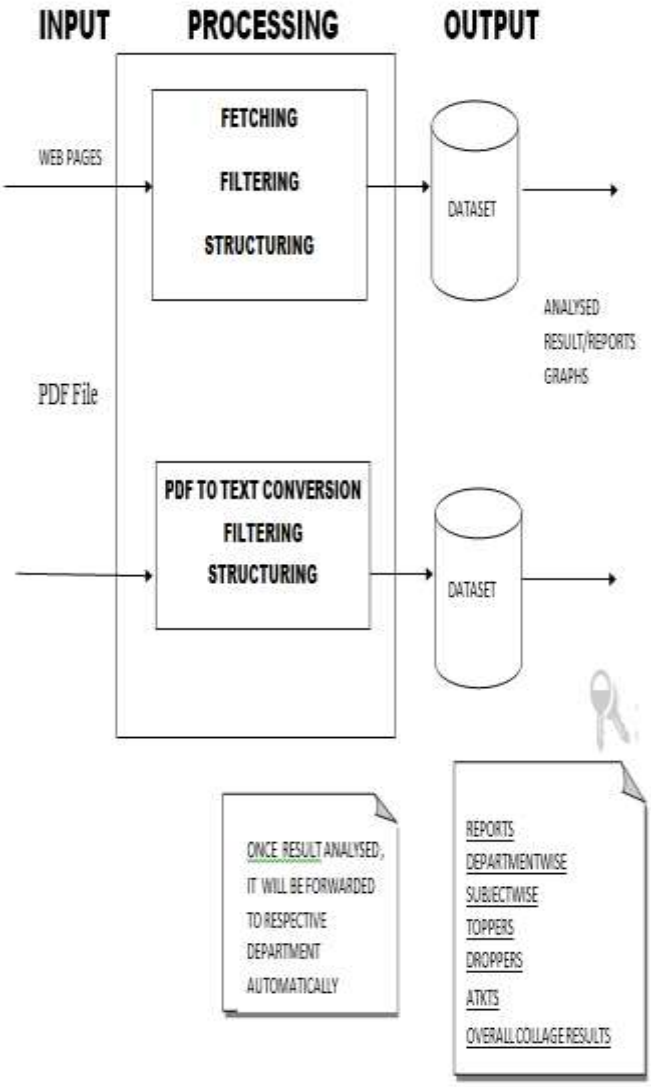

Fig. Basic Block Diagram

In this system, PDF file and Web pages are given as input to the system and generated reports are the output of the system.

Following diagramatic representation shows the detailed view of the proposed system :

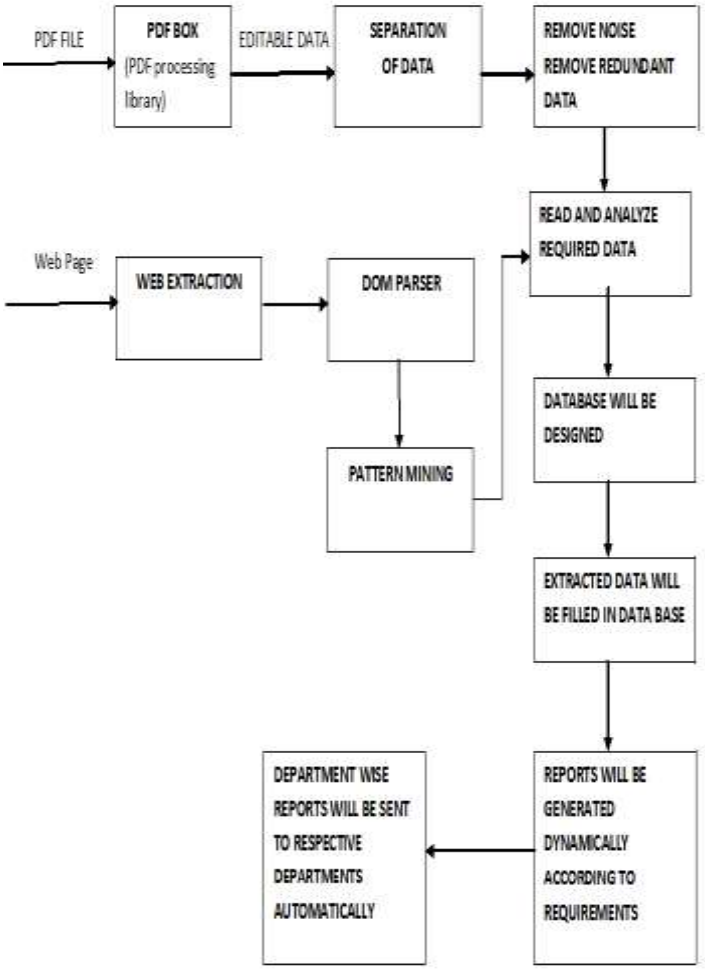

Fig. Detailed View

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 from Pune

University so it does not contain any diagram or images. To extract data from PDF files, we use PDF box technique.PDF box is actually PDF processing library.PDF box has ability to quickly and accurately extract text from PDF documents. To use PDF box technique, we have to include iTextSharp package. iText is used by .net, android, JAE, java developers to provide enhancement to their application with a PDF functionality. It provide feature like 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.

2. Separation of data :

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

3. Remove Noise Remove Redundant Data :

After separation of required data from the extracted PDF data, the data which is not required for processing is to be removed. For this purpose, line by line parsing is done. Also the PDF contain lots of redundant data E.g. PDF contain same subject list for each student for his/her respective department. Then such redundant data is also removed and only one copy of data is stored in the system.

4. WEB Extraction :

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

5. DOM Parser :

DOM is Document Object Model. System uses DOM parser to organize the nodes extracted from web pages into the tree structure.

6. Pattern Mining :

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

7. Read and Analyze required data :

After removing the noisy and redundant data, system has required actual data. Then this data is accessed for each student. Analysis of each student data is to be done by the system. It involves reading subject list of particular department, dividing subjects into theory, practical, term-work and oral wise. Also system read personal information of each student from text extracted from PDF.

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

All gathered data which is required and filtered need to be store into the system. Thus system designs database dynamically. After database is designed, department wise tables are generated. Then analyzed data is to be stored into the tables. Also student information is stored in the different table.

9. Reports generated:

Reports are generated using the data is stored in the database. The reports like department wise topper, subject wise topper, ATKT's, dropper student, etc. System provides the functionality to mail the generated reports to the respective departments.

IV. PLATFORM AND TOOLS

We used C#.net as our programming language.

We have made use of StarUML as modeling language to generate the use-case diagram, sequence diagram, timing diagram etc.

Also, for database management we have used Microsoft SQL and we used the QTP as our testing software.

V. EXPERIMENTAL RESULTS

We have conducted experiments to see how the reports are generated and how the analysis is done on different PDF and Web pages . The proposed system works faster on large PDF's with relative ease.

CONCLUSIONS

The proposed system automate the works to analyze the results and generate different reports and graphs as per user interest of user for Pune University and Mumbai University, thus reducing manual work and time .

ACKNOWLEDGEMENT

We express our sincere gratitude to Prof. Smita Patil (Assistant Professor, K.K.W.I.E.E.R.) for her support and guidance.

We would also like to thank Prof. Mrs Chitali Patil (Asst. Professor, KKWIEER) for her valuable words of advice.

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. Rendón-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, nimrod}@cenidet.edu.mx

[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.

[5] Figure Metadata Extraction From Digital Documents.Sagnik Ray Choudhury†, Prasenjit Mitra†‡, Andi Kirk_, Silvia Szep_, Donald Pellegrino_, Sue Jones_, C. Lee. Giles†‡†Information Sciences and Technology, ‡Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802 USA_The Dow Chemical Company, Spring House, PA 19477 USAsagnik@psu.edu, pmitra@ist.psu.edu,{andikirk,sszep,dapellegrino,susanjones}@dow.com, giles@ist.psu.edu