Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444
Volume 3, Issue 2, February-2016

SELF REGULATIG RESULT ANALYSIS FROM RAW DATA TO A PRESENTABLE FORMAT

Abhishek Pilania, Avinash Joshi, Abhijeet Kapade, Chandan Kumar
Dept. of Computer Science Engineering, Savitribai Phule Pune University.
G.H.R.C.E.M, Wagholi, Pune, Maharashtra, India.

Abstract — The result that an institution receives from the Savitribai Phule Pune University (SPPU) is in PDF (Portable Document File) format file. To analyze and manipulate them manually is very tedious job for a person. This is where text mining and file format convertors come into light to overcome this kind of problem. Text mining has an exceptional contribution for programmers and scientists to mine data out from textual files. Text mining overlooks the redundant textual data and only targets only on the required data patterns. Text mining plays important role because its center of functioning is devising patterns into necessary format or could be said that it derives a very quality of information from text. File format conversion has always been a bridge as well as challenge to access that particular data without particular tool. Different files have protocol set like internal and external meta-data, file headers, etc. as per application needs. So to form a bridge or so called convertors are designed as per our/system requirements by studying acutely of input file and the resulting or output file. So by adopting the technologies/ideas like text mining and file format convertors we are able of committing this project. It is an approach towards automation in educational field by using the data evaluation and manipulation

Keywords- *Text mining, PDF, JEE*

I. INTRODUCTION

The times of results publishing in newspapers are gone, today's world witnesses technology revolution that brings out a very dramatic change in lives of very people. Now is this era of technological revolution a bunch of people analyzing computer generated result manually is very unfortunate circumstance and in this circumstance there are chances of errors, biasing ,etc. . The result that an institution receives from the Savitribai Phule Pune University is in PDF format file. This particular project is adapting the concepts of file format convertor that converts the given PDF result file (i.e. the input file) into the text format file. The conversion of PDF file to Text file in J2EE (Java) the itext API is used. This text file acts as intermediate file, as data manipulation on text is when compared to PDF but when compared to database then anyway database is preferable. This is when the prominent and vital technology i.e. text mining comes into existence ,here the control goes word to word for pattern matching , and eliminating redundant data out of the converted text file(the intermediate file). By using preprocessors and report generators in Graphical format or Microsoft Excel (i.e. the output) are generated as per the user's needs. Reports are generated in Microsoft Excel formats also to correlate with previous year results.

II. LITERATURE SURVEY

1] Improving the Table Boundary Detection in PDFs by Fixing the Sequence Error of the Sparse Lines

The research paper that is published in 2009 10th international conference on document analysis and recognition by Ying Liu, Kun Bai, Prasenjit Mitra, C. Lee Giles College of Information Sciences and Technology The Penn State University, highlighted that PDF documents are very easy for converting in text format and that text data can be easily analyzed rather that converting the PDF file to a HTML format or image format by using techniques like OCR and the text data is more accurate than that of other formats.

But problem with this approach was that tools used which were used for conversion of PDF to text faces a problem of text sequence error for overcoming that problem two algorithms were used for recovering the sequence error and these algorithms were based on table boundary detection methods.

Algorithm 1sorting sparse lines across the document columns.

The improving procedure comprises two portions: the cross column Resorting, and the within-column resorting. Here the column denotes to the text column as an alternative of the table column. In what way to grow the text column evidence is out of possibility of this tabloid. We accept a relaxed but active method by devious the consistent length of the non-

sparse outlines then linking with the text measurement. For the inside-table script order disarranging, we only device the within-column resorting. Out beyond-table script arrangement disarranging, we have to contrivance together resorting's.

Algorithm 2: Sorting sparse lines within a document Column.
To contract with the overhead superior cases, we suggest additional text categorization algorithm without seeing the document column material at the start.

The procedure has four stages:

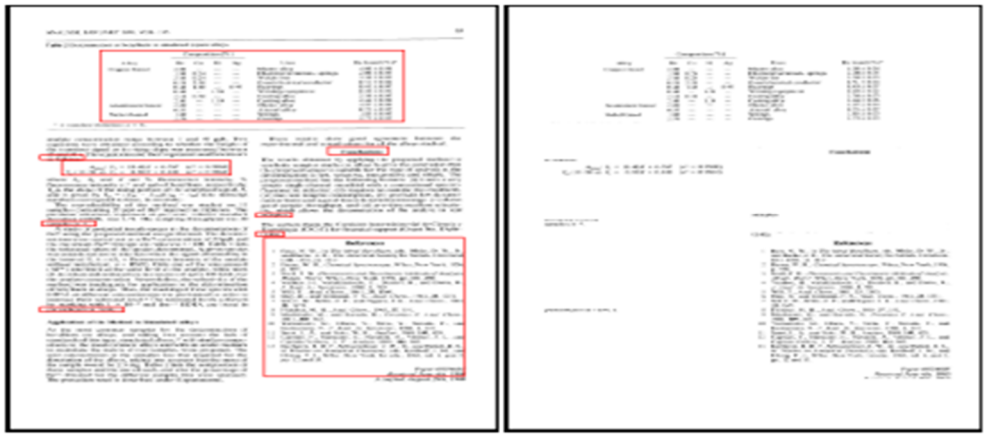
1) Gaining all the non-duplicated Y-axis values of the scarce lines within an area;

2) Categorization these Y-axis values from the top to the bottom;

3) Organization all the spare lines in the area conferring to the sorted Y-axis values;

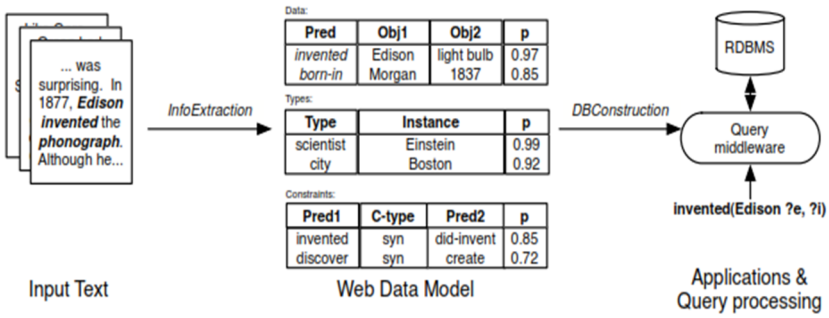
4) Spreading the sparse lines in the area arranged.

For the inside-table text arrangement error delinquent, the area is the table himself. For the beyond-table text arrangement error problem, this procedure workings improved for the extensive tables because the area refers to the entire page. All the spare lines of cross-column tables will be transferred out successively. The high-risk case of this algorithm is the parallel tables.



2] Structured Queries over Web Text

The study by Michael J. Cafarella, Oren Etzioni, Dan Suciu University of Washington stated that web is having is very vast amount of data and most of data is present in text format only and is unstructured data. But in the most of cases text data is also consist of structured elements and these structured elements requires text query processing the data in precise way.



3] Effective Pattern Discovery for Text Mining

There are lots of data mining methods which have been suggested for excavating important patterns in a text documents. Most current text mining approach simple centered approaches, they all undergo from the glitches of polysemy as well as synonymy. This paper offers an advanced and operative pattern discovery system which embraces the routes of pattern organizing with pattern progressing, for improving the efficiency of consuming and apprising discovered patterns for ruling applicable and motivating evidence.

III. DESIGN & IMPLEMENTATION

Here the design demonstrations who the system is working towards :

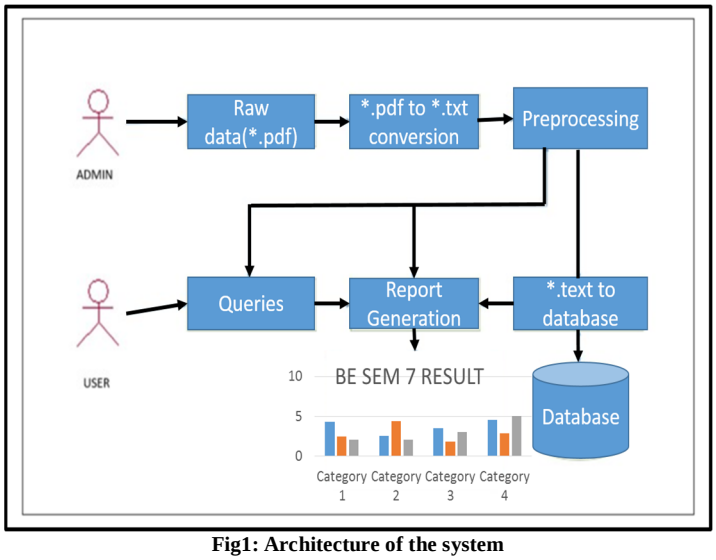


Fig1: Architecture of the system

- In this system client\user is only one is having who is having the actual interaction with the system by means of uploading the PDF file to system and one more small role is there for the user to fire the different quires.
- Admin accept the input as PDF file admin not any human being it's an automated controller for the system.

The system architecture is mainly consist of 3 modules:

1. PDF to Text
2. Preprocessing on text file
3. Text to Database
4. Database to reports

1. PDF to Text:

In this module PDF file is converted to a text format by using some of the java API one of the API is I-Text API which uses line by line text parsing and converts the pdf file to text format.

2. Preprocessing on text file:

Here in this module text format of the pdf is provided as input, preprocessor removes the redundant data. So we can say that preprocessor is a state which makes easier to handle data by applying some predefined rules to the text format.

3. Text to Database:
After preprocessing is done the preprocessed text is now stored in database by tuple format.

4. Database to Reports:
User fires quires for generation of reports with the help of the preprocessor the report generation module collects the data from the database and generates the different as per required by the user they may be in chart or table format.

IV. CONCULSION

There are numerous tools and software existing for extracting important information from PDF files especially for mining of text data but there are many problems with these available tools which are available online or openly. The main aim of drastic reduction in manual work and creating an automated system which takes a PDF file followed by generated report in presentable format using various vital technologies and approaches in which text mining and file format convertor had played crucial role in successful generation of desired reports in graphical and Microsoft Excel Files.

V. ACKNOWLEGMENT

We here by wish to take this opportunity to express our gratitude to our teachers and friends and all who have helped toward the completion of our project. We also like to give thanks to our Guide Mrs. Poonam Gupta for helping us and guiding us throughout our endeavor. We are very grateful to our teaching sta for guiding us all over the duration of the degree. They were very helpful to us, as and when we required their help. We are also very grateful to non-teaching sta to support us in the research laboratory in numerous ways.

REFERENECES

[1] *http://www.pdfparser.org/*
[2] *http://josefrichberg.squarespace.com/journal/2010/1/6/simple-series-ssis-importcolumn.html*
[3]Citeseerx,*ist.psu.edu/showciting?cid=180372*
[4]*Research.microsoft.com/en-us/projects/cryptanalysis/aesbc.pdf*
[5]*Crypto.stackexchange.com/related-key-attacks-on-aes.html*
[6] Hyubgun Lee, Kyounghwa Lee, Yongtae Shin, *AES Implementation and Performance Evaluation on 8-bit Microcontrollers, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6 No. 1, 2009.*
[7] Ritu Pahal, Vikas kumar, *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013*
[8] Julia Juremi Ramlan Mahmod, Salasiah Sulaiman Jazrin Ramli, Enhancing Advanced Encryption Standard S-Box Generation Based on Round Key , International Journal of Cyber-Security and Digital Forensics (IJCSDF) 1(3): 183-188 The Society of Digital Information and Wireless Communications (SDIWC) 2012 (ISSN: 2305-0012)