Credit Based Result Analysis System By Extracting Data From Pdf

[1]Bhushan Dashpute, [2]Vaskar Priyanka, [3]Jadhav Karina, [4]Ms. Yogita Desai

Dept. Of Information Technology, SNJB, Chandwad

SNJB's Late Sau. K.B. Jain College of Engineering, Chandwad, Dist. Nashik Maharashtra, India

Abstract

Current exam activities are mostly done on paper. Automated solution using this system covering the most important drawbacks of manual system, namely speed, and simplicity. The proposed system is focused on creating a Credit Based Result Analysis System by Extracting Data from PDF. The manual method of students' academic result processing was found to be tedious, especially when carried out for a large number of students, this makes the entire process time- consuming and error prone. The Portable Document File (PDF) format has become a popular means of producing documents in portable format. However, once a document is stored in this format, it may generally be considered to be" read only" and the PDF Reader software produces the analyzed data. The system extract data from PDF for that it uses di erent algorithms such as line picker, PDF parser, Boundary extractor. This system can be used for reading a PDF document as well as to extract the information in required format.

Keywords- Result analysis system, Extracting data, PDF, analyzed data, algorithms.

INTRODUCTION

Today's result declared by Universities for engineering is in PDF le format. In college level analysis of result done manuall y and it is time consuming and tedious, they are inaccurate or make many assumptions. 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 manuall y 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 proposed one system which would automate the process of result evaluation and analysis. This system takes 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.

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 LAPDF Text 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, there is complete details are discussed about data extraction.

The PDF-Box Technique: to extract references from PDF which converts the PDF data into text and get the require information from data. The PDF-Box technique is used for extracting references makes use of regular expressions. Besides that, classification techniques such as the K-NN algorithm, the Nave Bayes algorithm, Similarity of Cosine and Euclidian Distance are also employed here. The extraction process receives as input a PDF le generated from an original digital document (PDF-text le) and produces as

output the set of all references present in the paper, after the text Reference is first identified. For the extraction process two strategies are combined to yield better results.

The LA-PDF Text Technique: The LA-PDF Text technique which is a command line utility to extract text from PDF just by providing path of PDF le. The methodology created to extract the required information from the PDF les, classify and instantiate them in the document ontology.

PROPOSED SYSTEM

Following figure shows the detailed view of the proposed system:

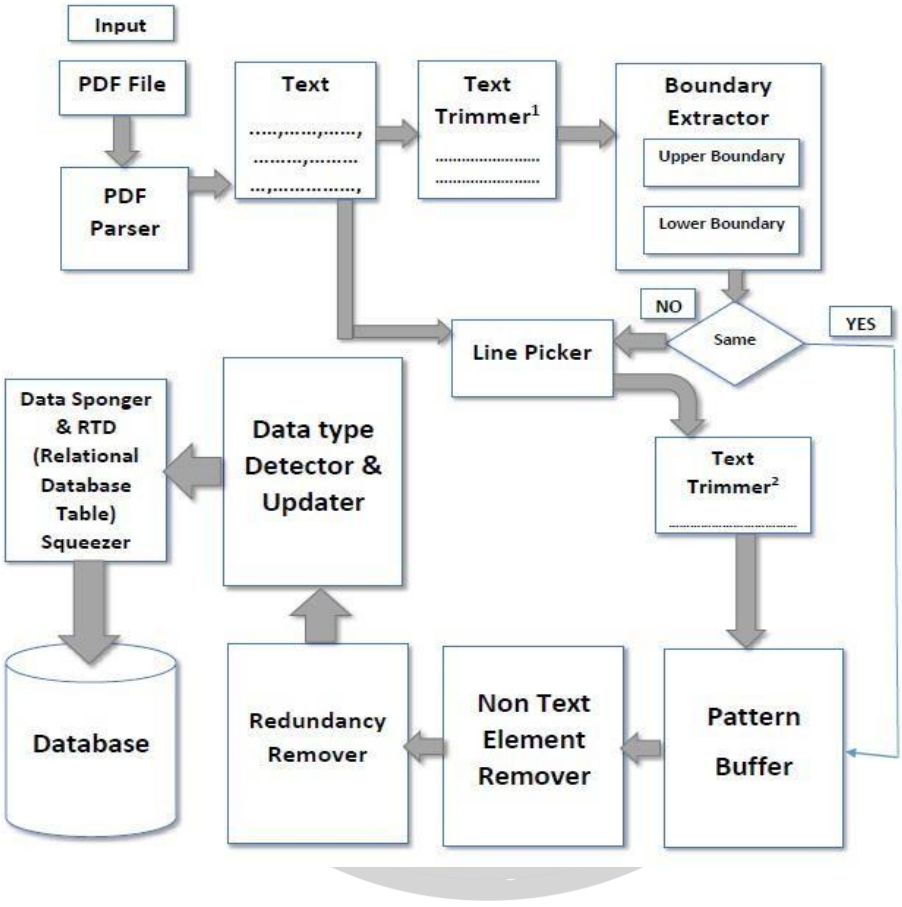

Fig., System Architecture

PDF Parser:

Pdf-parser is a command-line program that parses and analyses PDF documents. It provides features to extract raw data from PDF documents, like text, compressed images. The tool can also be used to extract data from damaged or corrupt PDF documents.

Text Trimmer:

Text Trimmer1 takes plain text copy of the extracted data and does the same thing that Comma Remover does but alongside it also removes white spaces. It first replaces all white spaces with commas and then removes all commas generated and those already present in the plain text. The same algorithm in the form of Text Trimmer2 is applied at the output of Line Picker.

Boundary Extractor:

After getting commas and white spaces removed from plain text (TEXT) form of data, the top row (first boundary) and the bottom row (second boundary) are picked up for later comparison from the other copy after processing.

Line Picker and Pattern Buffer:

Tag less copy of a source code is fed to the Line Picker. This algorithm parses the whole tag less copy. It starts with picking up first line of the file and provides it to the Text Trimmer which further converts it into a form comparable with the output boundary extractor.

Non-Text Element Remover:

The unformatted line from Pattern Buffer might contain non text html elements such as ” ”. This module removes all such pattern form pattern buffer and handover redundant file to redundant buffer

Redundancy Remover:

This algorithm checks and replaces multiple occurrences of commas with a single comma and marks it as a field separator. This process also takes care of the fact that there should not be any leading and trailing commas so it detects such commas and removes them from the file.

Data Type Detector and Updater:

This algorithm assumes that the first line in a file is the perfect candidate for the list of column headings in the relational table in the relational database and rest of the lines are data to be populated. Building on this assumption it starts with second line and checks for the data type of each value and if it is found to be of string type or date type it encloses the field value in quotes so as to avoid any data type mismatch error with SQL insert query. The output generated is the text file whose contents are guidelines and inputs to the target database table.

Data Sponger and RDT (Relational Database Table) Squeezer:

This algorithm acts as a sponger for the formatted textual data stored in a” valid data and squeezes the sponged data onto the table in a database. Before squeezing the data, it first creates a blank relational database and then a blank table inside the database taking the first row of” valid data file as a column heading and from algorithm and it forms a metadata of datatypes of the columns. Once blank relational table is generated it starts a loop from second line of sponged textual data, fetches it, forms an Insert SQL Query and then fires it. The result is populated table with squeezed data from a sponger.

REPORTS GENERATION

Reports are generated using the data is stored in the database. The result reports will be generated 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.

CONCLUSION

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). By this way result data will be organized well, which becomes easy to manage the result records.

ACKNOWLEDGMENT

We express our sincere gratitude to Prof. Ms. Yogita Desai (Assistant Professor, SNJB's KBJ COE) for his support and guidance. We would also like to thank Prof. Ms. Madhuri Kawade (Asst. Professor, SNJB's KBJ COE) for his valuable words of advice. We are also extremely grateful to our respected H.O.D. Dr. M. R. Sanghavi and Principal Dr. M. D. Kokate 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 Amazonas Manaus, Brazil Rafael Dueire Lins, Universidade Federal de Pernambuco Recife, Brazil Maria Lencastre, Universidade de Pernambuco Recife

[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 Sehga**l 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 files **FANG 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.

