International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

An Approach for Extracting Result Data from PDF Documents and its Classification

Ms. Jasmeen M. Maniyar[1], Ms. Pooja S. Pangarkar[2], Ms. Vaishnavi S. Kshatriya[3], Ms. Prerana N. Shinde[4] ,Mrs. Tejaswini A. Borhade[5]

[1] BE Student, Dept. of Computer Engineering, SVIT, Nashik, Maharashtra, India.

[2] BE Student, Dept. of Computer Engineering, SVIT, Nashik, Maharashtra, India.

[3] BE Student, Dept. of Computer Engineering, SVIT, Nashik, Maharashtra, India.

[4]BE Student, Dept. of Computer Engineering, SVIT, Nashik, Maharashtra, India.

[5]Professor, of Department (Internal Guide), Dept. of Computer Engineering,

SVIT, Nashik, Maharashtra, India

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.

Key Words*: *PDF Extractor, Pattern matching techniques, data mining, Web mining technique and sorting technique.

1. INTRODUCTION

SPPU has recently introduce a credit-based system to analysis the performance of student which was introduced in academic year 2015-2016. Result is based on SGPA and CGPA and Credit earned by student. Predicting difficulty level of subject for student is also provided by proposed system.

The examination result PDF is declared by the university and give to the individual colleges. This result file is in a single format and very hard to distributes in various departments and analysis toppers and fail students so this proposed system show the result which is calculated and create the separate form in distinction, first class, higher second class, second class, pass class, department wise toppers, subject wise toppers and also provide the functionality to predict the difficult subject in the next year based on current understanding level of the student. The system is developed for analysis of student result and generate the department result.

In this system, PDF file and Web pages are given as input to the system and generated reports are the output of the system. To develop a system which will extract the data automatically from PDF and web, create dynamic database and analyze data, as per the queries given. 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 PDFs with relative case.

The security can also be provided as per the requirements because only these systems handled by authorized person which is approved by admin. The main thing is that our system reduces the human works at a great effort.

It can be helpful such that

- Records are always updated.
- Manpower is decreased or reduced.
- Large amount of data regarding department and their modules can be stored.
- Accurate and perfect calculations are made.
- Maintenance of file is efficient and flexible.

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] With the tremendous amount of information that becomes available on the Web on a daily basis, the ability to quickly develop information agents has become a crucial problem. A vital component of any Web-based information agent is a set of wrappers that can extract the relevant data from semi structured information sources. In reference [2] This paper describes an approach for extracting information from PDF files. The key idea is to transform the text information parsed from PDF files into semi-structured information by injecting additional uniform tags. An extensible rule set is built on tags and another knowledge. Guided by the rules, one pattern matching algorithm based on a tree model is applied to obtain the necessary information. A further experiment proved that this method was effective. In reference[3] With the rapid increase of the PDF files in Internet, how to manage and search PDF files efficiently and quickly has become an urgent problem to be solved. The most important step of solving this problem is to extract information from the PDF files. This paper presents a new method for extracting information from PDF files. It first parses PDF files to get text and format information and injects tags into text information to transform it into semi-structured text, and finally, one pattern match algorithm based on tree model is applied to obtain the solution. A further experiment proved this method was effective. In reference [4] Presented a content metadata extraction (ME) framework from scientific articles using the case-based reasoning (CBR) method where they emphasized that the most important problem of content extraction from web sources is classification of HTML tag sequences. They argue that accurate classification of HTML tag sequences which contain scientific metadata provide valid content extraction of a scientific paper represented in web documents. The next component, pattern extraction, classifies HTML tag sequences as either informative or non-informative patterns. In the last step, content extraction, extracts contents of the academic article from the classified patterns of the HTML tag sequences. In CBR approach, they used HTML and its features to extract contents of a scientific paper.

3. PURPOSE

Result analysis requires large amount of manual work. Our system works for Pune university Engineering colleges And Mumbai University Diploma Colleges results. In 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 Pune university, automatically stores the data into the database, once the database is created, we can extract various information from that data using various queries.

4. PROPOSED SYSTEM

Following figure shows the System Architecture of the proposed system:

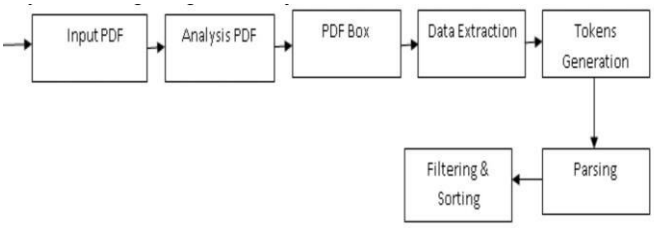

Fig -1: System Architecture

4.1 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 provides 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.

4.2 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.

4.3 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.

4.4 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

4.5 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 departments, dividing subjects into theory, practical, term-work and oral wise. Also, system read personal information of each student from text extracted from PDF. After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

5. MATHEMATICAL MODEL

System Description:
S = (I,O,F)
Where,
S: System.
I = { PB,SD RN, PM} are set of Inputs
Where,
PB : PDF Box
SD : Separation of data
MS : Remove Noise
PM :Pattern Mining

F = {F1, F2, F3, F4}are set of Function
Where,
F1 : PDF Extraction
F2 : Pattern matching
F3 : Web Data Extraction
F4 : Pattern Mining

O = {O1} are set of Output
Where,
O1: Alert User

Success Conditions:
Proper inputs will provide proper output.
Failure Conditions:
No database, internet connection

Web Mining Algorithm

Algorithm for non-terminal nodes N

- Initialize

  textCnt(N) = 0 and linkCnt(N) = 0

  set S(N) as an empty set.

  setT extCnt(N) = 0 and setLinkCnt(N) = 0

- For every child of N,

  textCnt(N) = textCnt(N) + textCnt(child)

  linkCnt(N) = linkCnt(N) + linkCnt(child)

- Calculate

  childRatio = (textCnt- linkCnt)/ textCnt

  If childRatio > Threshold

  add the child to S(N).

  setTextCnt(N)=setTextCnt(N)+textCnt(child)

  setLinkCnt(N)=setLinkCnt(N)+linkCnt(child)

  Store S(N), textCnt(N), linkCnt(N), setT

extCnt(N)

and setLinkCnt(N)

6. CONCLUSIONS AND FUTURE WORK

This system is been published on website and pdf files are been sent to college, College staff has to manually find results and have to make rankings accordingly. It will sort all the data according to their marks, for this we will use Web mining techniques, PDF Extraction, data fetching and sorting techniques, which will make user to simplify the data easily and make results accordingly. The goal of the system is achieved and difficulties are solved. The project is built such that it is user friendly. Analysis of the scoring system it shows by the grade wise result of individual subject and final result also display grade wise. depending on its range of marks. The project can be easily used in college for college result analysis of student. It reduces time which required for manual calculation.

In Future We will provide an Android application for the same working. Make the system generalize for any type of result format and auto learner. Previously, data used to be

inserted manually to analyze result. But, Currently the project supports excel(.xlsx) files for extraction of data. The future scope is that data can be fetched, parsed in other formats like doc, csv , odt , etc .

ACKNOWLEDGEMENT

We express our sincere gratitude to Prof. Mrs. Tejaswini A. Borhade (Professor, SVIT) for his support and guidance. We are also extremely grateful to our respected H.O.D. Mr. Kishor N. Shedge and Principal Dr. G. B. Shinde 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] Muslea, Steve Minton, and Craig A. Knoblock. A hierarchical approach to wrapper induction, Proceedings of the Third International Conference on Autonomous Agents, Seattle, WA1999, 221-227 M.Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] Zhu Ming, Wang Jun, Wang Junpu. Multiple Record Extraction from HTML Page Based On Hierarchical Pattern, Computer Engineering, 2001, 27(9): 40-42.

[3] A New Method of Information Extraction From PDF FILES FANG YUAN1, 2, BO LIU.

[4] Prasath. R. and Ozturk. P. "An Approach to ContentExtraction from Sciatic Articles using Case Based Reasoning". Research in Computing Science 117(2016). pp.85–96.

BIOGRAPHIES

Name: Jasmeen M. Maniyar.

Educational Details:
B.E.Computer (Pursing.)

Name: Pooja S. Pangarkar.

Educational Details:
B.E.Computer (Pursing.)

Name: Vaishnavi S. Kshatriya.

Educational Details:
B.E.Computer (Pursing.)

Name: Prerana N. Shinde.

Educational Details:
B.E.Computer (Pursing.)