Volume 2, Issue 1, January 2017 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

A Strategy for Automatically Extracting References from PDF Documents

Akshay Prakash Chavan
Computer Engineering
Sanghavi College of Engineering
Nashik, India

Tejas Shriram Bhalerao
Computer Engineering
Sanghavi College of Engineering
Nashik, India

Niraj Jitendra Chhajed
Computer Engineering
Sanghavi College of Engineering
Nashik, India

Tejas Nitin Nagare
Computer Engineering
Sanghavi College of Engineering
Nashik, India

Prof. P. Biswas
HOD &Asst.Professor
Computer Engineering
Sanghavi College of Engineering
Nashik, India

*Abstract: **Every day the number of citations an author receives is becoming more important than the size of his list of publications. The automatic extraction of bibliographic references in scientific articles is still a difficult problem in Document Engineering, even if the document is originally in digital form. This paper presents a strategy for extracting references of scientific documents in PDF format. The scheme proposed was validated in Live Memory platform, developed to generate digital libraries of proceedings of technical events. document processing, regular expression, learning. 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

The acknowledgement of the sources of a technical article is in its list of bibliographical references. Conversely, the number of citations a given article receives may be an indication of its importance in a given area. Thus, citation indices are becoming more important than the size of the list of publications of a given author or researcher. Collecting such information is far from being a trivial task, however. In the case of legated paper documents an effort of paramount dimension is necessary. This is due to the need to either re-type such data or to scan the document, to automatically process it, in order to enhance the quality of the image, attempt to find the list of references, and finally transcribe it via OCR . Such scheme is still processing intensive and error prone. In the case of electronically generated documents of formats such as PDF, PS, HTML and XML, the task of reference spotting is much easier, and tends to be more accurate than in the case of legated printed ones. This does not mean that it is a straightforward task. The automatic extraction of references is still a difficult problem in Document Engineering. In proceedings, neither authors use, nor editors check to guarantee that the adopted bibliographic templates were strictly followed. Problems often arise in the items in the list of references, such as: incompleteness, existence of different formats out of the pattern, abbreviations, etc. This work details

a process for extracting bibliographical references in the context of the LiveMemory Project . LiveMemory is a platform developed for the semi-automatic generation of digital libraries of proceedings of technical events. It allows processing the image of scanned documents (JPEG, TIFF e PNG), automatic indexation of files, extraction and storage of information in databases, such as: paper title, authors and their institutions, keywords, abstracts and references, year of publication, etc. The platform also allows the generation of reports about most used keywords, most cited references, etc. This paper presents the strategy used in the Live Memory platform for extracting the list of references of a PDF document, which was digitally generated. Regular expressions, together with classification and identification based on K-NN algorithm and the Naïve Bayes algorithm were used for this purpose. The whole process is presented throughout this paper, which is organized in the following way: Section 2 presents related work in the literature; Section 3 details the strategy for extracting references; Section 4, presents the extraction system, which gives support for the proposal; Section 5 presents some performance tests; Finally, Section 6 presents the conclusions and draws lines for future work. Result analysis requires large amount of manual work.

Our system works for 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. LITRATURE SURVEY

Several researchers proposed ways of extracting information from bibliographical references. This section describes some of such work and also tools that are close to our proposal. The first work on bibliographic reference extraction used the Hidden Markov Model (HMM) technique . In reference the authors consider the tagging process for classifying the items that compose references, and also the automatic induction of a set of rules for extracting specific features. Reference extracts information from texts in Japanese using OCR. First, blocks are labeled with title, abstract and references; after each block is re-labeled for the extraction of information that one requires. In reference the authors propose the extraction of the names of authors from academic papers, using the identification of uppercase letters, lines breaks, tagging of characters and use of regular expressions. Aljaber and his colleagues use the scope of the citations to verify the similarity between texts and the partition into classes for applying the K-Means algorithm. In a combination of regular expressions, a system based on heuristics and knowledge is proposed. In a system was developed for extracting information from texts containing scientific citations; they consider a hybrid approach based on automatic learning, which combines text classification techniques with the Hidden Markov Model (HMM).

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

III. SYSTEM ARCHITECTURE

Basic block diagram of proposed system as 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 diagrammatic representation show

Architectural Design

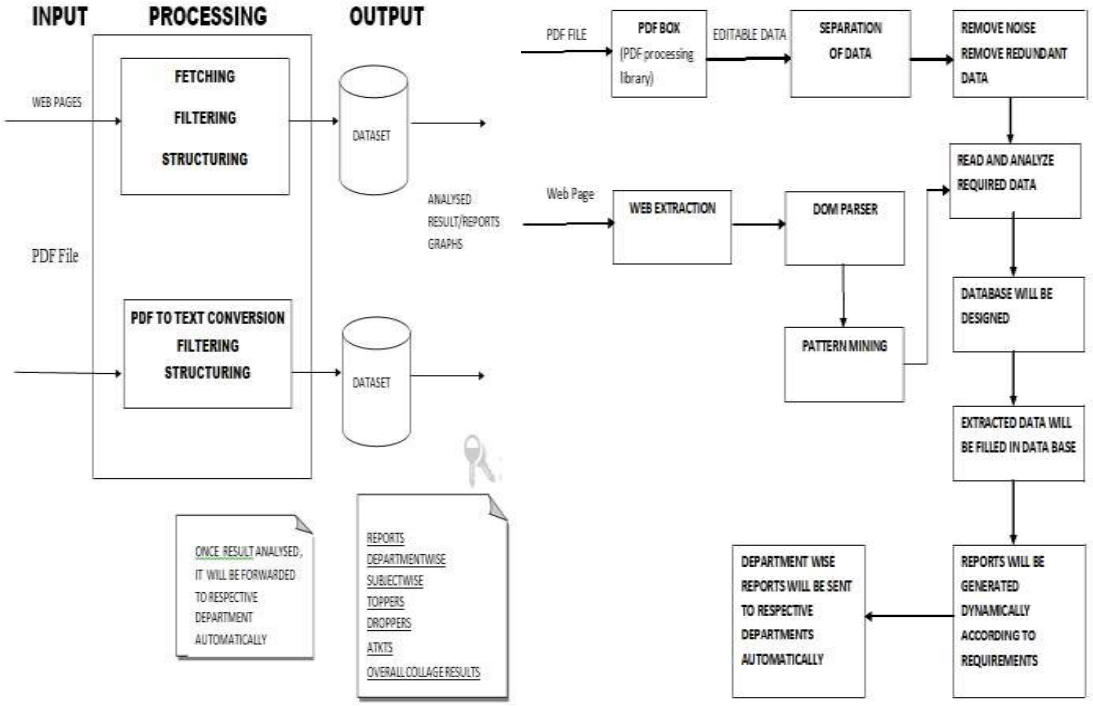

Figure:1 Block Diagram

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 PuneUniversity 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.

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.

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.

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.

DOM Parser:

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

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 specified pattern and mined the data accordingly.

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.

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.

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.

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.

EXTRACTION TOOL

This section describes the extraction system, which gives support to the proposal. The system has a specific interface for the **Training Phase**, which guides the following steps:

• The user inputs the references to be analyzed;

• The system partitions each supplied reference and find the punctuation marks: point, comma and semicolon;

• Each partition or fragment is analyzed by the system and fills in the corresponding vector of characteristics;

• Then, the user classifies each fragment, selecting its label as: title, author or other information;

• Finally, the system analyzes the training base generating general vectors with the measures obtained for each classification: title, author.

Figure2 shows the training phase and the accuracy obtained from the process. The test phase is similar to the training phase, but the classification elements are automatically chosen by the system, and the vectors values are calculated. For each fragment, the system compares the calculated values with the General Vectors obtained. For the comparison phase three strategies are adopted. The first one to be adopted is the K-NN algorithm with Euclidean distance, which is used to verify the similarity between the vector of fragments that is being considered with the general

vector. The same is done using the K-NN algorithm with Cosine Similarity and finally the Naïve Bayes algorithm was tested. After making some case studies using the 3 approaches, it was observed that K-NN algorithm with Euclidean Distance and Cosine Similarity present equivalent performance, that is, both achieved improvement in title extraction, but part of the authors still remain classified as "other information", which is not satisfactory. Tests using the Naïve Bayes algorithm, which considers the likelihood preceding element, showed better results. Figure 4 shows diagram the Test Phase in block

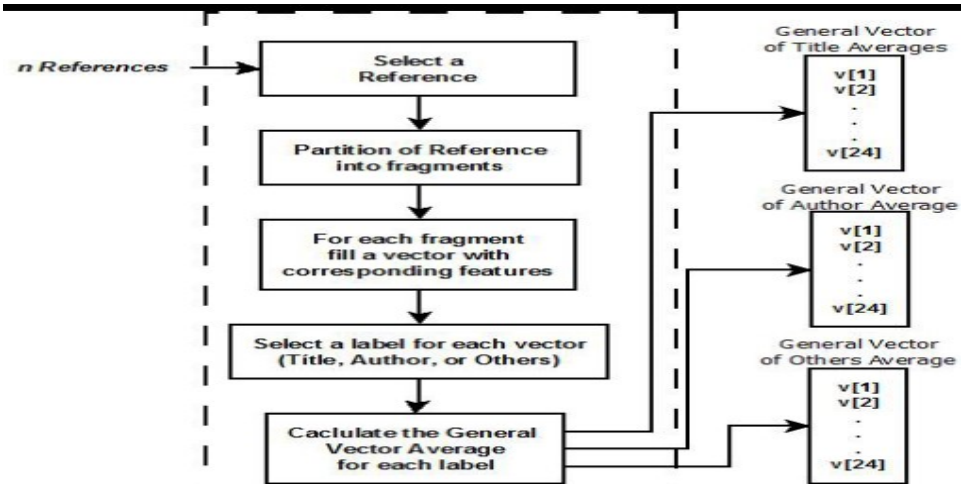

Figure 2. Training phase and extraction media.

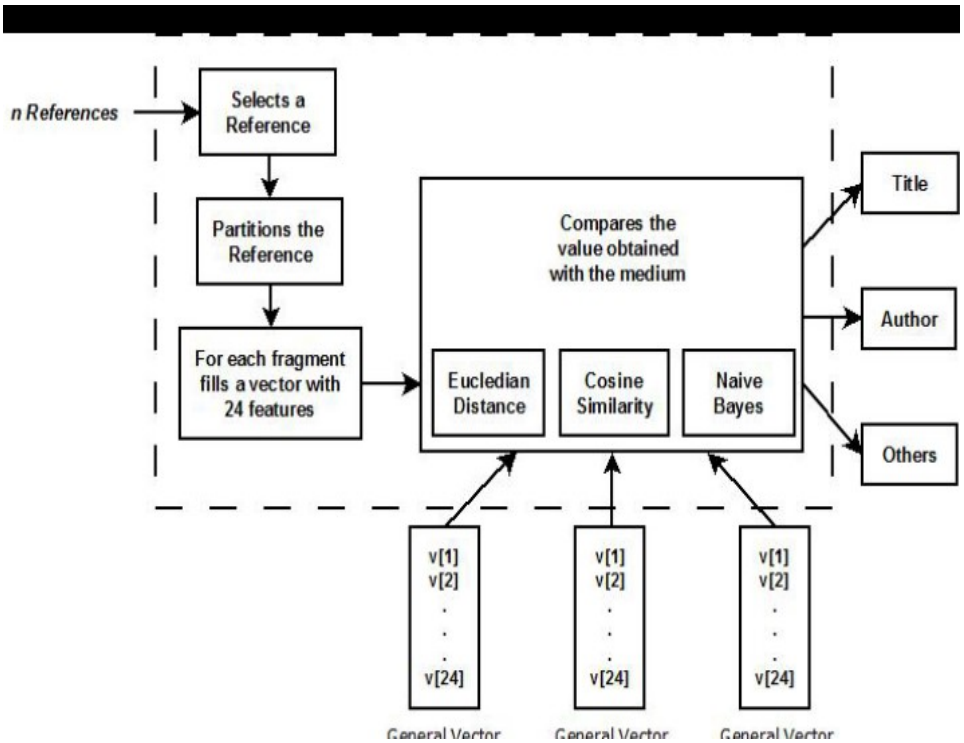

Figure 3-:.Shows the training interface.

CONCLUSIONS

The proposed systems 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.

REFERENCE

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
3. Cuernavaca, México {juancarlos, juliaarana, gabriel, nimrod}@cenidet.edu.mx
4. 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
5. 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.
6. Figure Metadata Extraction From Digital Documents.Sagnik Ray Choudhury, Prasenjit Mitra, Andi Kirk_, Silvia Szep_, Donald Pellegrino_, Sue Jones_, C. Lee. GilesInformation 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,dapellegrin o,susanjones}@dow.com, giles@ist.psu.edu
7. Abbyy FineReader Home Page. http://finereader.abbyy.com/.
8. Álvarez, Alberto Cáceres. Extração de informação de artigos científicos: uma abordagem baseada em indução de regras de etiquetagem.USP; 2007. Available at: http://www.teses.usp.br/teses/disponiveis /55/55134/tde-21062007-144352/.
9. Bader Aljaber; Nicola Stokes; James Bailey; Jian Pei. "Document clustering of scientific texts using citation contexts," Information Retrieval, V.13, N.2, 101-131, DOI: 10.1007/s10791-009-9108-x, 2009.
10. Constans, Pere. "A Simple Extraction Procedure for Bibliographical Author Field," Word Journal OF The International Linguistic Association, February, 2009, Available at http://arxiv.org/abs/0902.0755.
11. Gupta, D.; Morris, B.; Catapano, T.; Sautter, G. "A New Approach towards Bibliographic Reference Identification, Parsing and Inline Citation Matching," In Proceedings of IC3. 2009, 93-102.
12. Hua Yang; Norikazu Onda; Massaki Kashimura; Shinji Ozawa. Extraction of bibliography information based on image of book cover. In Proceedings of the 10th International Conference on Image Analysis and Processing IEEE Computer Society Washington, DC, USA, 1999.
13. Ohta, M., Yakushi, T, Takasu, A. "Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields" In Proceedings of ICDIM, 2008, 99-104.
14. PDF-Box Home Page. Extracted from http://www.pdfbox.org, March 21 2011.
15. R. D. Lins, G. Torreão, G. F. P. e Silva. Content Recognition and Indexing in the LiveMemory Platform GREC 2009. Springer Verlag. LNCS 6020. p.220-230, 2010.
16. Silva, Eduardo Fraga do Amaral e. Um sistema para extração de informação em referências bibliográficas baseado em aprendizagem de máquina CIn-UFPE; 2004.
17. Tesseract Home Page. Extracted from http://code.google.com/p/tesseract-ocr/downloads/list, June 22 2011.
18. Atsuhiro Takasu, "Bibliographic Attribute Extraction from Erroneous References Based on a Statistical Model," jcdl, pp.49, Third ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'03), 2003.