Volume 2, Issue 5, May 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

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

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

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

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

Prof. P. Biswas
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 Live Memory Project .

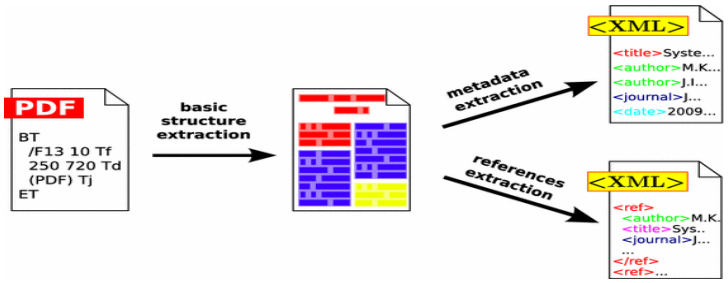


Fig:1 Extraction workflow architecture.

At the beginning, the basic structure is extracted from the PDF file. Then, metadata and bibliography are extracted in two parallel paths 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. RELATED WORK

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).

Table 1:extraction workflow into independent processing paths and steps

Path	Step	Goal	Implementation
A. Basic structure extraction	A1. Character extraction	Extracting individual characters along with their page coordinates and dimensions from the input PDF file	iText library
	A2. Page segmentation	Constructing the document's geometric hierarchical structure containing (from the top level) pages, zones, lines, words and characters, along with their page coordinates and dimensions	Enhanced Docstrum
	A3. Reading order resolving	Determining the reading order for all structure elements	Bottom-up heuristic-based
	A4. Initial zone classification	Classifying the document's zones into four main categories: *metadata*, *body*, *references* and *other*	SVM
B. Metadata extraction	B1. Metadata zone classification	Classifying the document's zones into specific metadata classes	SVM
	B2. Metadata extraction	Extracting atomic metadata information from labelled zones	Simple rule-based
C. Bibliography extraction	C1. Reference strings extraction	Dividing the content of *references* zones into individual reference strings	K-means clustering
	C2. Reference parsing	Extracting metadata information from references strings	CRF

III. PREPROCESSING ONLINE DOCUMENTS

A. *Preprocessing Online Documents* :

One problem with automatic reference linking is that not all formats (bitmaps, TeX, PDF, PostScript, etc.) are equally easy to parse. For this reason before an online document is analyzed, the first step is usually to transform the document into a format more susceptible to analysis. The two most common target formats are ascii and xhtml (the xml version of html). ResearchIndex (formerly CiteSeer)uses a version of pstotext that inserts font tags into the document as the document is converted from PostScript/PDF into ascii. 3 Similar approaches are used in analyzing OCR conversions from bit maps. Summers derives paper segments (such as title and authors) from inspecting the geometric layout of a scanned document. More recently, Caton pointed out that presentation directives can be used to generate tags that help navigate a document. In general, these various formats with their font notations can be converted to html tags, and then analyzed by our XHTMLAnalyzer software. The software from Southampton[10], which reference-links PDF files found in the arXiv repository at Los Alamos, uses Acrobat tools to convert PDF into ascii text prior to analysis. Likewise, discusses the preprocessing of Word documents into a form that can be analyzed. In general, the conversion tools listed in Table 1 are recommended for preprocessing full-text documents into a form that can be analyzed.

Full-text format	Conversion algorithm	Analyzable format	Layout info
ASCII	no conversion	ASCII	none
HTML	Tidy/JTidy	XHTML	HTTP tags
DVI	dvips, pstoascii	ASCII	fonts
PostScript	pstoascii	ASCII	fonts
PDF	pdfps, pstoascii	ASCII	fonts
bitmaps	OCR	ASCII	? depends
Word	save as HTML, Tidy/JTidy	XHTML	HTTP tags

Table 2: Conversion tools to prepare for parsing.

A. Extracting an Items Metadata

Why is it important to have an item's bibliographic data when analyzing it for reference linking applications? The main reason is that since we are analyzing this item, we have the online location of this item. It is a linkable copy of a work. In order to know what work that is, we need to know the item's bibliographic data, such as title and authors and year of publication. Once we have determined the work of which this item is a copy, then if we come across this work in a reference list in the future, we already know that we have a linkable reference, and we know its location. Either the item is accompanied by metadata, as is the case with Open Archive items and more recent D-Lib papers, or else it has to be extracted from the text of the paper. We do the latter. To extract the metadata for an analyzed item, layout clues are necessary. Usually presentation information (such as font changes) is used to determine what the title is. Titles usually occur in a large font, near the beginning of a paper. In html, one can assume the title is contained in an or element, although it happens sometimes that the title is simply set off by a element that increases the text size. Multiline titles can be extracted from html documents by reconciling the parsed title.

Metadata Extraction Algorithm (for XHTML)

settitle1 = value of <title> element if there is one
Scan for any of the following:
<H1>text</H1>, <H2>text</H2>, text,
text, text
settitle2 = "text"
if title2 is shorter than title1,
then scan for subtitle and append to title2

IV. 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

A.Architectural Design:

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 :

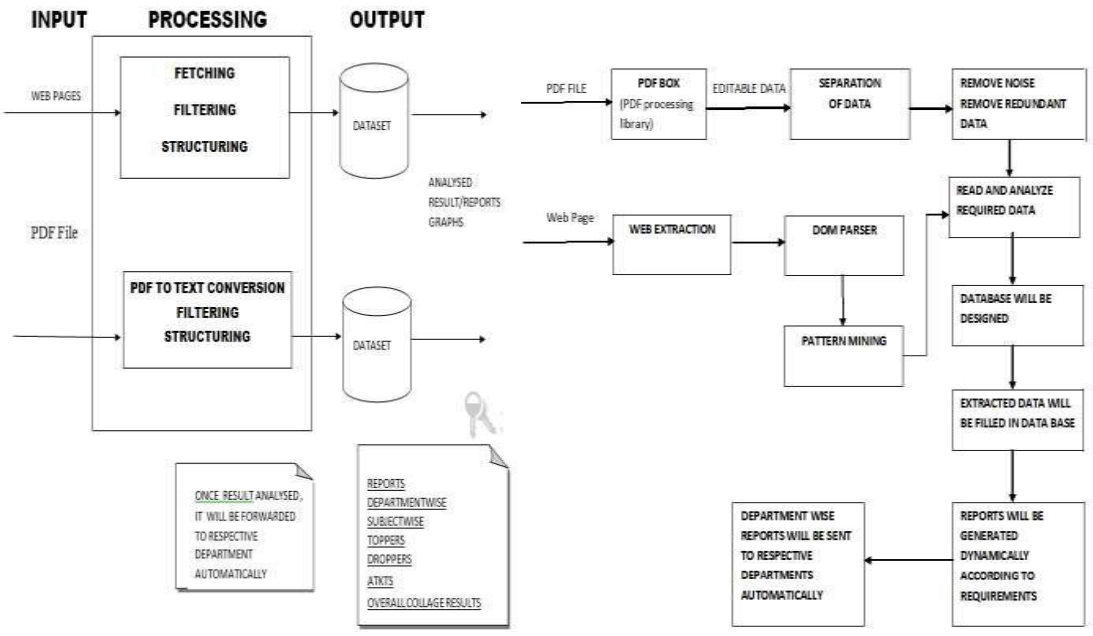


Fig:2 Block Diagram

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.

B. 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.

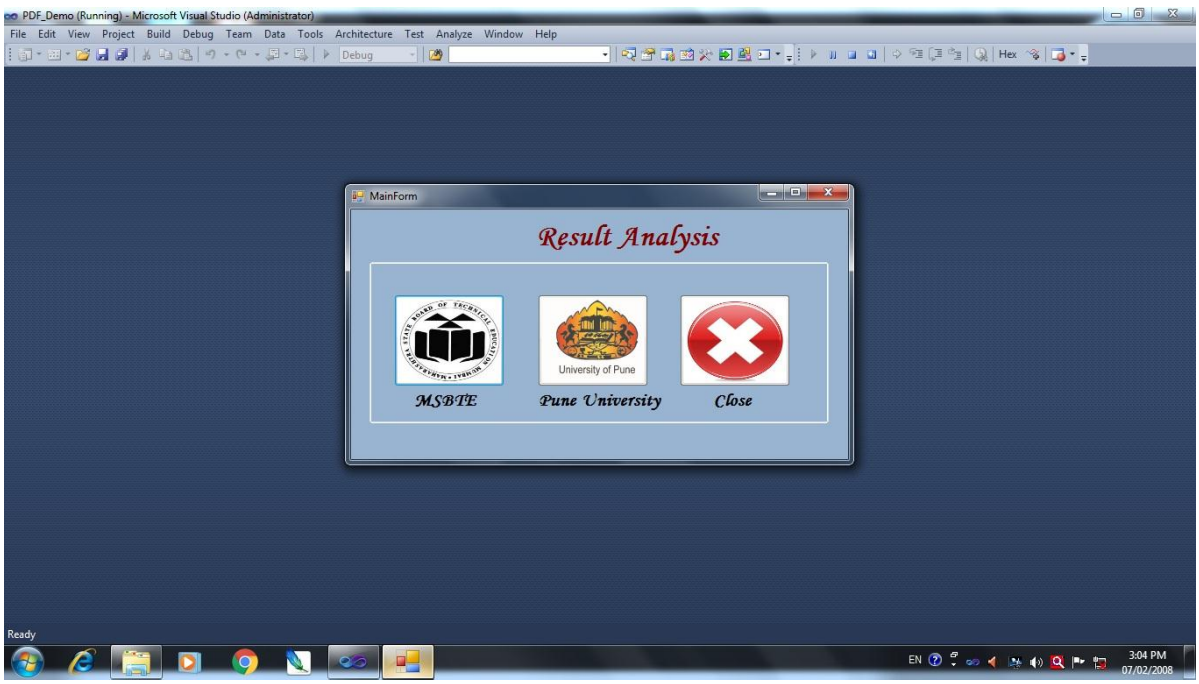
Pattern Mining :System uses pattern mining method to find the required data from extracted document. The extracted plain text by the web extractor is checked this the specified pattern and mined the data accordingly.

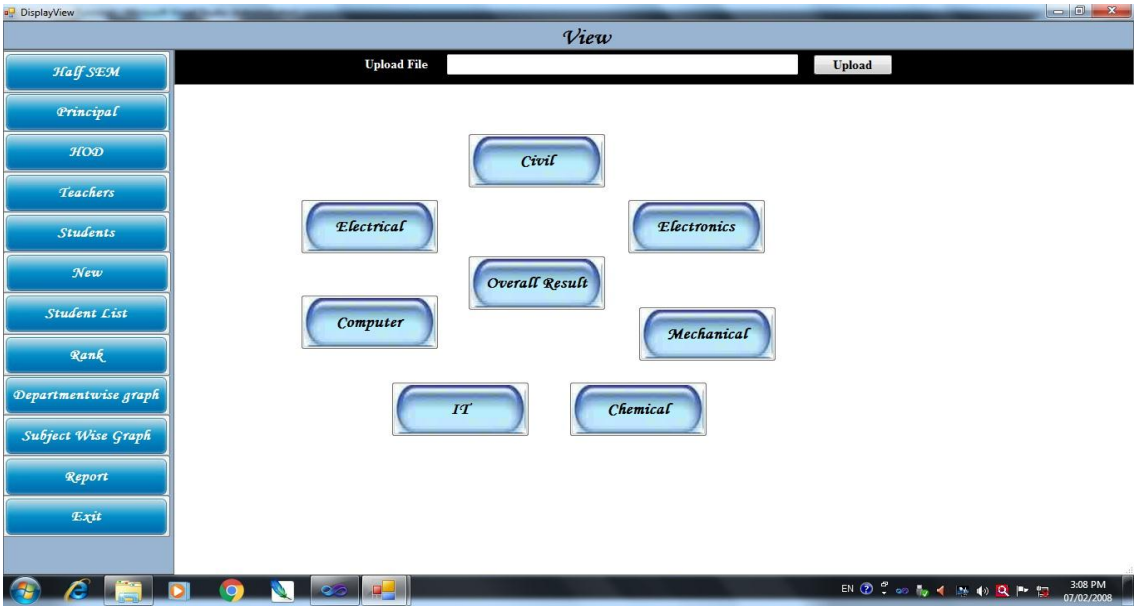
Read and Analyze required Data :After removing the noisy and redundant data, system has required actual data. Then this data is accessed for each student. Analysis of each student data is to be done by the system. It involves reading subject list of particular department, dividing subjects into theory,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

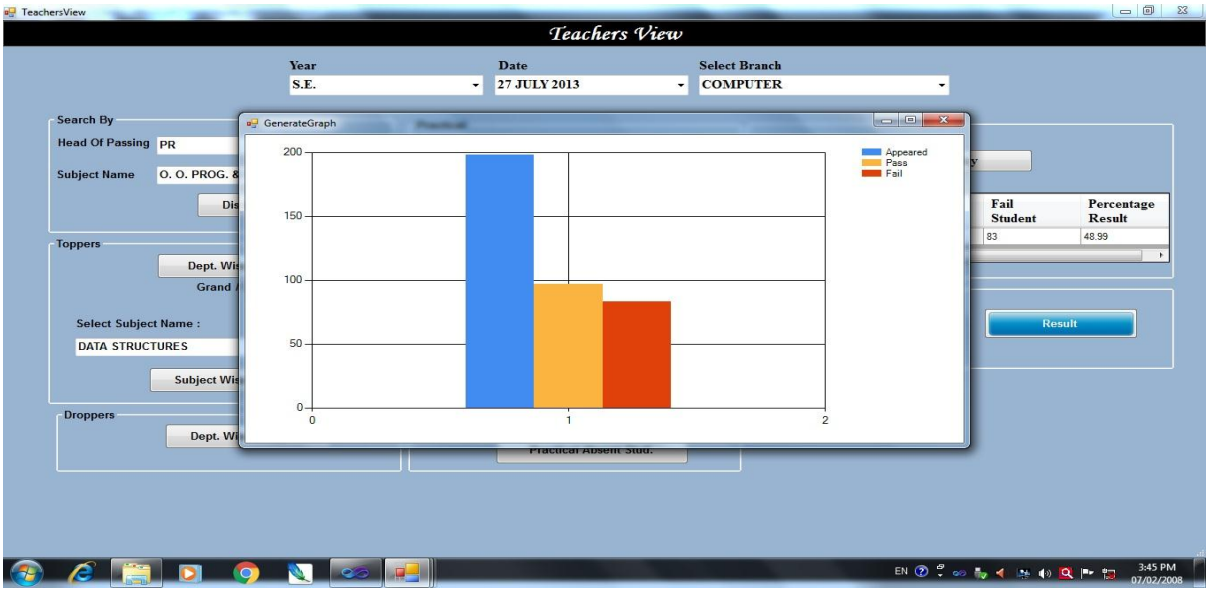
Database designed and extracted data filled in the system :All gathered data which is required and filtered need to be store into the system. Thus system designs database dynamically. After database is designed, department wise tables are generated. Then analyzed data is to be stored into the tables.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

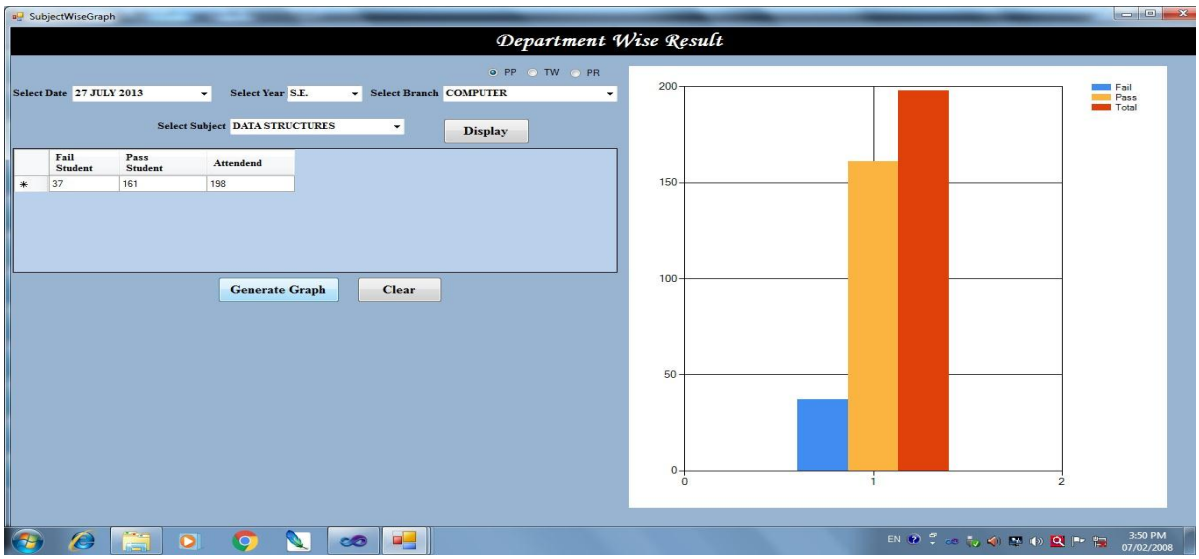
Reports generated:Reports are generated using the data is stored in the database. The reports like department wise topper, subject wise topper, ATKT' s, dropper student, etc. System provides the functionality to mail the generated reports to the respective departments.

V. SCREENSHOTS




DisplayView
View
Half SEM
Principal
HOD
Teachers
Students
New
Student List
Rank
Departmentwise graph
Subject Wise Graph
Report
Exit
Upload File
Upload
Civil
Electrical
Electronics
Overall Result
Computer
Mechanical
IT
Chemical


TeachersView
Teachers View
Year
S.E.
Date
27 JULY 2013
Select Branch
COMPUTER
Search By
Head Of Passing
PR
Subject Name
Toppers
Select Subject Name :
DATA STRUCTURES
Droppers
GenerateGraph
Appeared
Pass
Fail
Fail Student
Percentage Result
83
48.99
Result


SubjectWiseGraph
Department Wise Result
PP
TW
PR
Select Date
27 JULY 2013
Select Year
S.E.
Select Branch
COMPUTER
Select Subject
DATA STRUCTURES
Display
Fail Student
Pass Student
Attendend
37
161
198
Generate Graph
Clear
Fail
Pass
Total

VI. RESULTS

In this section we will detail a specific study made to validate the proposal made in this paper.

A. Training

In the Training Phase, 16 papers were used, and from those 98 references were extracted and partitioned in 892 fragments, of which 363 fragments were classified as “titles”, 417 as “authors” and 112 fragments were classified as “others”. The feature vector, described in Table 1, shows how these fragments were classified. One may observe that there are some cases where the feature is specific of a field, such as the “year”, this is practically only given for the field “others”. If the title is in the database, the “other” fields are verified there also, eliminating the possibility of redundancy in the stored data, considering that there are references which have difference only in the year of publication.

B. Results

For validating the reference extraction scheme proposed here, the SBrT database was used [9], specifically the papers published in 2010, the papers are in English, because the event was the ITS’2010 - IEEE/SBrT International Telecommunications Symposium. Papers were in text editable PDF format. At first, the focus was in the identification of references, and the strategy was to use only 10 papers from each year, which include 186 bibliographic references. When the regular expressions were applied to the 10 papers, the following results were obtained:

• 117 “titles” were correctly identified, representing 62,90% of the existing tiles;

• 101 “authors” were identified, that is, 54,30% of the total;

• 170 “years” were identified, that is, 91,40% of the total;

• 122 were identified as “others”, and the accuracy rate was 65.59% of the total. Table 2 shows the results when regular expressions are applied together with the Naïve Bayes algorithms. The classification of each reference element was:

• For “titles”, 138 were correctly identified, yielding 74,19% accuracy;

• For “authors”, the system correctly identified 114 instances, providing an accuracy of 61,29%;

• For “years”, 173 were correctly identified, that is 93,01%;

• For the “others”, 128 were identified corresponding to 68,82% accuracy

TABLE II. Results

Field	Regular Expression (RE)		Naive Bayes and RE	
	Accuracy	(%)	Accuracy	(%)
Title	117	62,90	138	74,19
Author	101	54,30	114	61,29
Year	170	91,40	173	93,01
Others	122	65,59	128	68,82

Figure 8 also presents the same results, but in a graphic way. From the graphic it is clear that the one where the

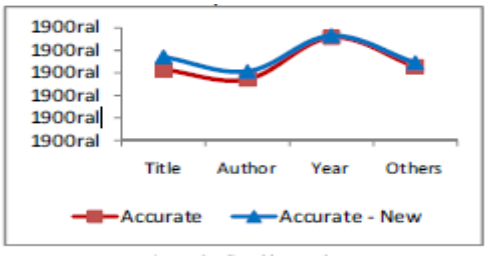


Fig.8. Graphic Results

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.

REFERENCE

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 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. 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
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.
19. 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.
20. 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/.
21. 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.
22. 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.
23. 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.
24. Silva, Eduardo Fraga do Amaral e. Um sistema para extração de informação em referências bibliográficas baseado em aprendizagem de máquina CIn-UFPE; 2004.
25. 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.