International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-1 Issue-5, June 2012

Chapter extraction from research documents using Meta-Content Framework

Tripti Sharma, Sarang Pitale

Abstract*—Automatic chapter extraction from electronic documents has always been an interesting task for researchers who are continuously engaged in subjective answering systems. Researchers are agreed on the fact that chapter extraction is one of the key processes to generate the model answers. The proposed paper presents a framework to extract the chapter contents from the research documents. The framework is implemented using Java technology and iText library , It takes research document of PDF format as an input and extracts the chapter contents in simple HTML format so that it can be easily rendered in web browser.

Index Terms—*PDF, Java, iText, Html.*

I. INTRODUCTION

Model answer generation is one of the interested domains for researchers of Natural Language Processing field. Imagine a digital library which generates answer to you, from one or more books, any where any time. Such a library can solve the problems of students belongs from rural areas or from remote places. To achieve such a task a framework is required which will process the e-document and give all the contents. A standard e-book typically contains various sections like front index, back index, list of tables, figures etc. the present paper shows a Meta-content Framework which can extract such sections so that the model answers can be generated.

Meta-content framework takes an e-book in PDF [1] format and applies a series of process to it for generating various content information, the processes in block diagram is shown in fig.1

A. File Splitter

This module splits the e-book of PDF format into number of pages which the book contains. For the splitting purpose a java [2] based library iText [3] is used which splits the PDF file into number of pages it contains. The process begins with the creation of a directory named exactly the same as of the file. After creating the directory the Meta Content framework splits the input file into number of pages and saves them into the directory. The naming convention of the book is as "book name-Page-No.-pdf". Finally, after completing the split process the inputs were ready for the next process. The split process is important because it provides a unique identification of the page and its respective contents.

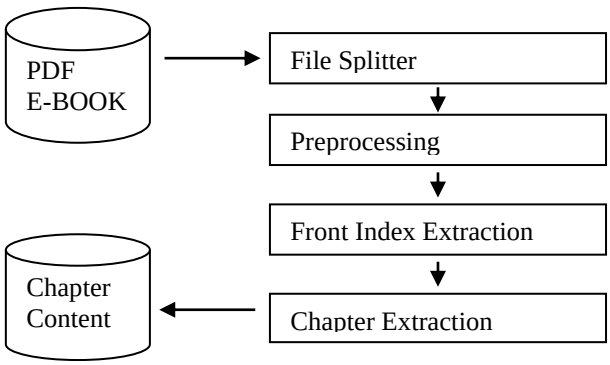

Figure 1. Block Diagram for chapter extraction

B. Preprocessing

After the successful split process, all the PDF pages are preprocessed for the conversion process. The preprocessing step is combination of three steps:

1) *Image Segmentation*
2) *File management*
3) *Text Conversion*

Image Segmentation is one of the crucial steps in Meta Content framework. This step detects the co-ordinate of an image in a PDF page. After getting the coordinates the schemes calculates its height and width, creates an html tag and write the tag in PDF file with the same name.

In between the segmentation process the scheme continuously deletes the old file and writes the new file until al the images had segmented. This task is done by File management step.

Text conversion is the last step of preprocessing. Various PDF to Text extraction tools [4] are available which have certain advantages and disadvantages. In this step a java based library, iText, is used for the extraction of text contents from the PDF files.

iText is used by César García-Osorio et.al.[5] For developing A Tool for Teaching LL and LR Parsing Algorithms .

C. Front Index Extraction

Front index of an e-book is generally table of contents part. Front index is a very useful factor for generating chapter contents. Front index contains chapter headers, section headers with page numbers, so a simple technique is to map those page numbers with the page numbers of e-book and extract all chapter contents.

Manuscript published on 30 June 2012.
* Correspondence Author (s)
Tripti Sharma, Department of Computer Science, Chhatrapati Shivaji Institute of Technology, Durg, INDIA.
Sarang Pitale, Department of Information Technology, Bhilai Institute of Technology, Durg, INDIA,

Simple analysis shows that front index of most of the books lies between first page and page number 15 to 20. Also the starting of front index page starts from the headings like "Contents" or "Table of Contents" and ends with "Bibliography" or "References", the present scheme uses the same facts and extracts the pages with these patterns. Standard documents like e-thesis, follows a standard and has a fixed heading for the front index and hence easy to extract front index.

Algorithm 1. Front_Index_Extraction(P_name,no_pag)
Input: P_name is a String which denotes a path of folder which contains the pages of e-book and no_pages denote the total number of pages of book.
Output: Pages containing front index
Step1: for i=1 to 20
each page in P_name
repeat
Step2: check if page content starts with "Table of contents" or "Contents" Mark page as start page of front index. Break and go to step 3
Step3: Check if page contains "Bibliography" or "References" Mark page as end of front index..
Step 4: extract start to end marked front index pages
Step5: return extracted pages as front index.
Step 6: end

D. Chapter Extractor

Output text files generated by the preprocessing module are now traversed to select those pages which are contributed as main chapters. This module extracts those pages which are the major contributing factors for back index generation. It is observed that the back index is mainly comprises of nouns from the chapters not from the references, acknowledgement etc. hence the module uses the same fact and separate those pages which are the part of chapters.

To perform such a task Meta content frame work uses a supervised String matching algorithm, which generally searches those pages which indicates the start and end of the chapter contents. After the pages were found, they are separated and saved in a directory named as chapter. Algorithm 2 shows the steps used to extract chapter content of a research document.

Algorithm 2. Chapter_Extraction(P_name,no_pag)
Input: P_name is a String which denotes a path of folder hich contains the pages of e-book and no_pages denote the total number of pages of book.
Output: Pages containing chapter
Step1: start= Front_Index_Extraction(P_name,no_pag)
Step2: Check if page contains "Bibliography" or References" after the front index pages. Mark page as end of chapter
Step 3: extract start to end marked pages
Step4: return extracted pages as chapter contents
Step 5: end

E. User Interface for framework

The Meta content frame work is developed using Java Technology. The master user interface is shown in figure2.The master user interface comprises of various menus which are listed below:

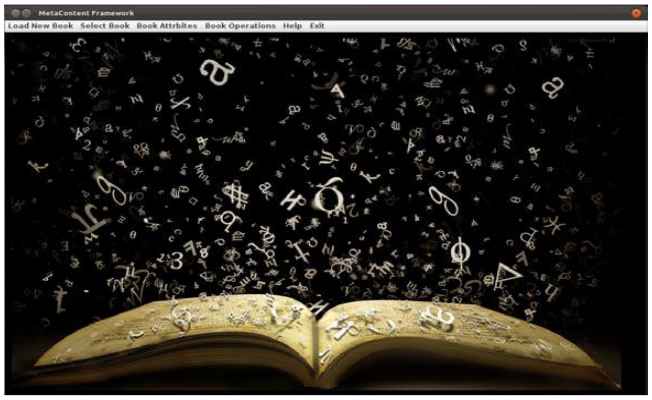

Figure 2. Front Index for meta content framework

1) Load Book:

This operation is used to load a PDF e-book to the frame work. After selecting this option a file-chooser helps the user to select a book from the drives, as shown in figure 3.

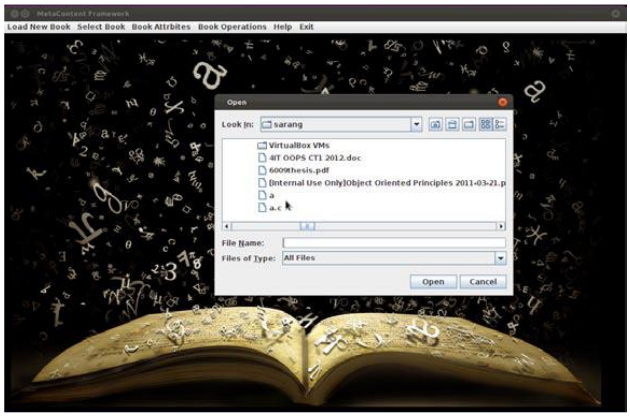

Figure 3. Book Loading using meta content framework

2) View attributes:

This operation lets user to view different attributes of the selected e-book. Various attributes of an e-book are:

- Name
- Size
- Number of pages
- Encryption status
- Rebuild status

The user interface to view attributes of an e-book is shown in figure 4.

3) Book operations:

This option lets the user to perform various extraction and generation based operations on e-books. Such as :

- Front Index Extraction
- Back Index Extraction
- Back index Generation
- Sectioning

The interface for this option is shown in figure 5.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

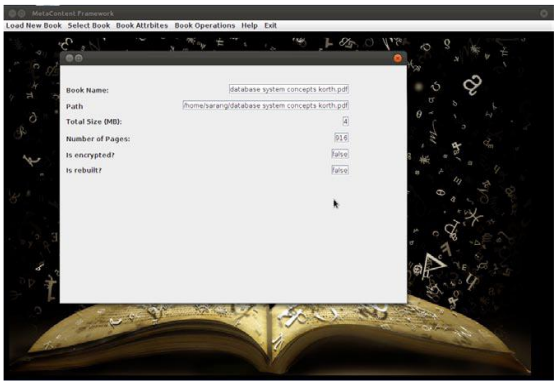

Figure 4. Book attributes

4) *Help and Exit:*

These are the general options which are same as the other available software's.

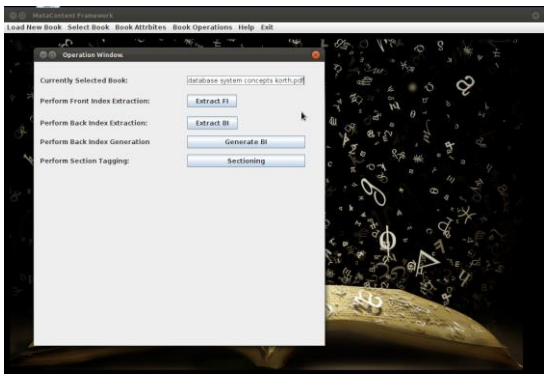

Figure 5. Book operations

II. META INFORMATION FROM META CONTENT FRAMEWORK

Proposed scheme uses e-thesis in PDF format as an input. PDF format e-thesis in Meta content framework is processed in series of steps as shown in figure6.

At last Meta content framework generates three useful meta-information, Front index, back index and chapter contents. This meta-information is intern used for many applications like topic spotting, model answer generation, topic relevancy computations and many more. A revolutionary and very useful utilization of this Meta information is in implementation of digital libraries.

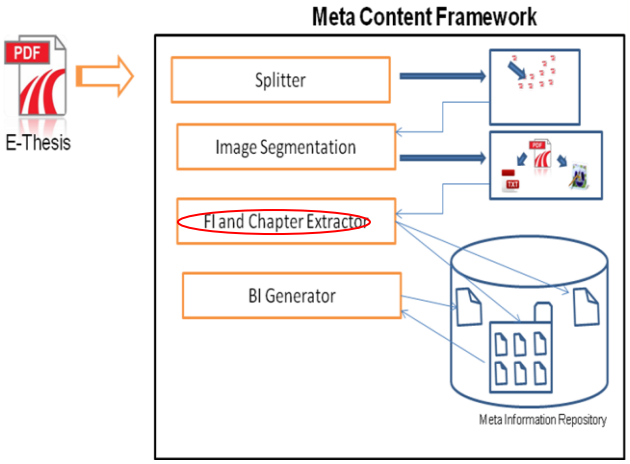

Figure 6. Working of meta content framework

As shown in figure 6 an e-thesis , a pdf format thesis is given for the processing to meta-content framework,the framework generates the three main components and chapter content is one of them.

III. RESULTS AND DISCUSSION

Several series of experiments were run to process e-books and it is observed that the time required for processing an input file depends upon the number of pages present in the book. The experimental setup was implemented upon text corpora from varied course ware (e-books in the present setup) and also with the different page counts and content sizes.

All the observations clearly indicate that the predominantly prevailing factor that contributes to excessive execution time is the varying document page size.

IV. ACKNOWLEDGMENT

The author sincerely thanks Prof. Arpana Rawal and Prof. Ani Thomas for their timely, invaluable and indispensable guidance and consequently encouragement shown towards the work group.

REFERENCES

[1] Adobe Systems Incorporated, http://www.adobe.com/pdf/

[2] Oracle Corporation, http://www.java.com/en/

[3] iText ®- Free / Open Source PDF Library for Java and C#, http://www.itextpdf.com/

[4] Sarang Pitale and Tripti sharma, "Information Extraction tools for portable document format", International journal of computer technology and applications,Vol 2 (6), 2047-2051

[5] César García-Osorio,Carlos Gómez-Palacios,Nicolás García-Pedrajas, "A Tool for Teaching LL and LR Parsing Algorithms", Proceedings of the 13th annual conference on Innovation and technology in computer science education, ACM New York, NY, USA ©2008, pp-317-317 M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science1989.

Tripti Sharma is Associate Professor in Department of Computer Science & Engineering at Chhatrapati Shivaji Institute of Technology, Durg (C.G.) , India. She pursued Bachelor of Engineering from University Institute of Technology, Bhopal and Masters Degree from Rungta College of Engineering & Technology, Bhilai. She is heading the affairs of Department of Computer Science Chhatrapati Shivaji Institute of Technology, Durg (C.G.) since 2010.Her research mainly focuses on image processing. She is a life time member of Indian Society of Technical Education, India (ISTE) and Institutional Member of Computer Society of India.

Sarang Pitale is Assistant Professor in the Department of Information Technology, Bhilai Institute of Technology, Durg (C.G.), India. He pursued his Bachelor of Engineering from Pt. Ravi Shankar Shukla University Raipur, Chhattisgarh. Currently, pursuing Master of technology, from Chhattisgarh Swami Vivekanand Technical Unversity, Bhilai. Sarang Pitale's primary research interests focus on Data Mining techniques. He is a member of IEEE, International Association of Computer Science & Information Technology, International Association of Engineers and Association for Computing Machinery (ACM).