Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 2, February 2015, pg.314 – 318

RESEARCH ARTICLE

EXTRACT AND ANALYSIS OF SEMI STRUCTURED DATA FROM WEBSITES AND DOCUMENTS

Prashant M Ahire[1], Anil P Gagare[2], Yogesh B Pawar[3], Savan S Vidhate[4]

[1,2,3,4] B.E Computer ,Department of Computer Engineering, GES's R H Sapat College of Engineering, Nashik 422005, Maharashtra, India

[1] ahireprashant77@gmail.com, [2] anilgagare840@gmail.com,

[3] yogesh.pawar95@gmail.com, [4] sa1.vidhate@gmail.com

Abstract:

Discovering into the W3 consortium and portable documents for the purpose of fetching useful information is a hectic task under the limitations of current available browsers. While a huge amount of work is being carried out to improve the efficiency. The huge amount of information on web and portable digital document is stored in backend databases which are not indexed by traditional search engines such databases are known as Semi structured Databases and extraction and analysis of web content and documents is a time consuming and complex task. Hence, there has been increased interest in retrieval and integration of semi structured web data and digital document data with a view to improve quality information to the users who wish to analyze the data. This paper states an approach that identifies web page templates and the tag structures of a portable document in order to sort semi structured data from web pages and documents and analyze the fetching data as per user requirement using various SQL queries.

Keywords: ***Web page extraction, Analysis, Web Page Service, portable documents***

I. Introduction

Data mining is nothing but the process of extracting useful information from collected databases. Extraction of the information from the big databases is called the "Knowledge Discovery". It is an analytical tool for analyzing data. It allows user to analyze data from many different aspects or angles, categorize it, and conclude the relationships identified. Technically, it is the process of finding correlations or patterns among loads of field in large relational databases [7].

Web mining is similar to the data mining where the data is extracted from the web pages. It is one of the applications of data mining techniques to discover patterns from the web. Web mining can be divided into three different types which are Web Usage Mining, Web Content Mining and Web Structure Mining. Web Usage Mining is the process of extracting useful information from server logs. E.g. use Web usage mining is the process of finding out what users are

looking for on the Internet. Web Structure Mining is the process of using graph theory to analyze the node and connection structure of a web site. Web Content Mining is the mining, extraction and integration of output oriented data, information and knowledge from web page content [3].

Extraction and analysis of the web pages and portable document is a heated research area in the field of data mining and web mining. Extraction is nothing but the fetching relevant information from the web page and portable documents. Extraction is crucial step for analyzing the data. Different web sites contain information on varied topics in various formats. Large amount of effort are often required for a user to manually locate and extract data of interest from the web pages and portable documents. Just consider about results of MSBTE University, the results are stored in HTML page format. If an analysis is to be made then gathering information, converting it into excel sheet to use various query to process the data resulting into subject wise result, toppers of each subject, overall topper etc. In the same way user will do analysis of Portable Document Format for that great efforts are needed [5].

II. System

2.1 Existing System:

Many extraction techniques have been reported for analysis of semi-structured database. Data Extraction Using DOM Tree and Selectors **(**DEUDS**)** a page level data extraction system that automatically discovers extraction pattern from web pages for selected data section and extracts data. The existing system uses a novel technique to extract table data from structured databases [1]. The first class of methods is based on machine learning, which requires human labeling of many examples from each web site that one is interested in extracting data from. The process is time consuming due to the large number of sites and pages on the web. The second class of algorithms is based on automatic pattern discovery. These methods are either inaccurate or make many assumptions. A plenty of work has been done in this area. Liu and Grossman proposed a novel method to mine data records in a web page automatically which is called as MDR. The technique is based on two observations about data records on the Web and a string matching algorithm. The technique of MDR is able to mine both contiguous and non-contiguous data records [2]. Its experimental results show that the technique outperforms existing techniques substantially.

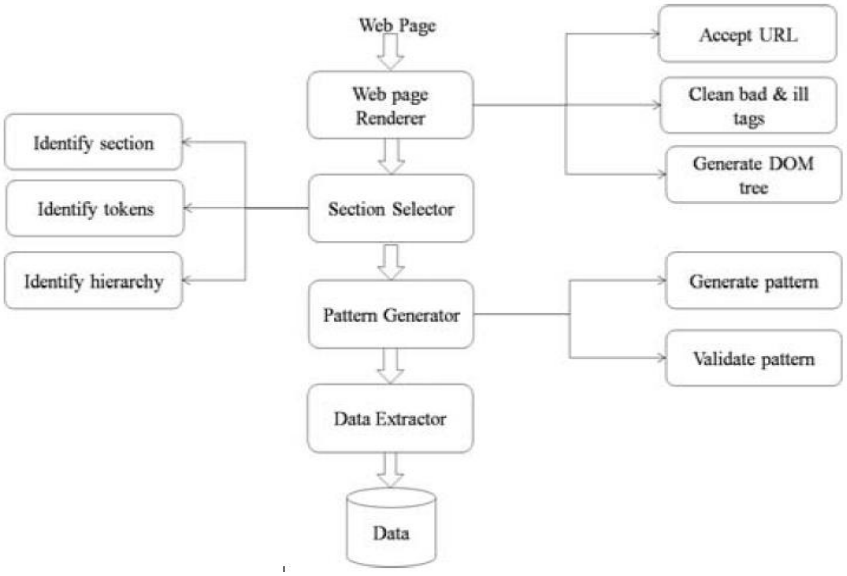

Fig: 3.1 Existing system

2.2 System:

Most of the work done has various problems like it takes a lot of time to extract; they are inaccurate or make many assumptions. The user will do analysis of web page or digital document for getting the output in easier way. In the proposed system the user select input as web page or digital document to which he/she want to extract and analyze the data. Then the proposed system uses different algorithms like line picker, web page renderer, pattern generator, etc to extract the data. The system puts important contents from web page or digital document into the database. The proposed system also uses the regular expression logic to match the important content to acquire the accuracy [3].

III. Methods

3.1.1 Web page extraction:

The main aim of the system is to have a data which is simpler to read to user. For that system is going to first extract the necessary data into database and then using different queries it produces the analyzed data. The system extracts data from web pages for that it uses different algorithms like line picker, boundary extractor, pattern generator, etc [4][5].

Algorithm for web page data extraction:

Input: Web pages.

Output: Extracted structured data in database.

1. Identify web page of which analysis is to be made.
2. Get HTML response of that web page.
3. Divide data using HTML tags.
4. For each line remove the HTML tags.
5. **Boundary extractor**: Remove header and footer contents which are not necessary.
6. **Pattern generator:** Match the structure with rules.
7. Extract text mode design schema in database.
8. Using different queries data to be analyzed is presented in tabular and graphical format.

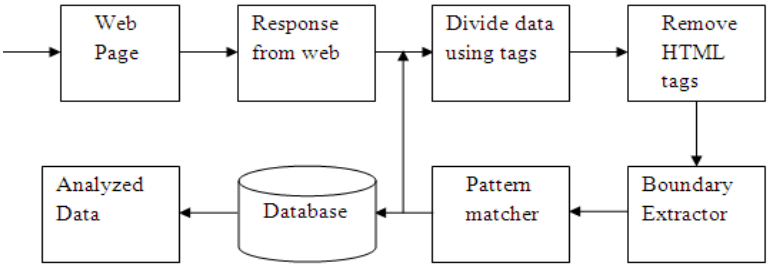

Figure: Web page data extraction

3.1.2 Digital document extraction:

Above algorithm is used to extract data from web pages in the same way the system extracts data from documents. Some of the digital documents contain semi-structured data so our system also extract and analyze such documents for that it uses algorithms such as line picker, PDF parser [8].

Algorithm for digital document extraction:

Input: PDF file.

Output: Extracted structured data in database.

1. Take the portable document as input.
2. Using **PDF parser** the document is converted into text file.
3. **Line picker**: Scan the text file line by line.
4. **Boundary extractor**: Remove header and footer contents.
5. Gather fields using regular expression and pattern matching.
6. Insert knowledgeable data in database.

7. Goto 3rd step.
8. Using different queries data to be analyzed is presented in tabular and graphical format.

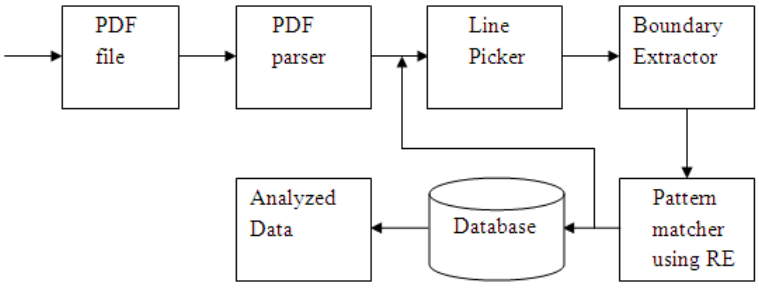

Figure: PDF data extraction

IV. Conclusion

In this paper, we implemented a new vision to gather semi-structured data from web pages. The problem is well known and is been performed by renowned researchers, of which existing techniques are either approximate or make many loads of assumptions. Main intention is to emphasize accurate & absolute result analysis in a speculated time in order to reduce rigorous calculations.

V. Acknowledgement

The satisfaction that accompanies the successful completion of any task would be incomplete without mentioning the people who made it possible. We are greatful to a number of individual's who's professional guidance along with encouragement have made it very pleasant endeavor to undertake this project. We have a great pleasure in presenting the project **"Extraction and Analysis of semi-structured data from web pages and documents"** under the guidance of **Prof. R. C. Samant** we are truly indebted and greatful to **Head of Department Prof. N.V. Alone** for their valuable guidance and encouragement. We would also like to thanks the Gokhale Education Society's R. H. Sapat, College of engineering, Management Studies and Research, Nashik-05 for providing the required facilities, Internet access and important books. At last we must express our sincere heartfelt gratitude to all the teaching and non-teaching members of computer department who help us for their valuable time, support, comments, suggestions and persuasion.

References

[1] Vinayak B. Kadam , Ganesh K. Pakle. "DEUDS: Data Extraction Using DOM Tree and Selectors." Published in International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1403-1410

[2] Manpreet singh sehgal and Anuradha. "HWPDE:Novel Approach For Data Extraction From Structured Web Pages " Published in International Journal of Computer Applications (0975-8887) Volume 50-No. 8 July 2012 pages 22-27

[3] Anuradha, A.K Sharma. "A Novel Technique for data extraction From Hidden Web Databases" Published in International Journal of Computer Applications (0975-8887) Volume 15-No. 4 February 2011 pages 45-48

[4] Teena Merin Thomas,V.Vidhya. "A Novel Approach for Automatic Data Extraction from Heterogeneous Web Pages" Published in International Conference on Emerging Technology Trends on Advanced Engineering Research (ICETT'12)

[5] A. Niranjana, Dr. V.Vidhya "A NovelApproach for Automatic Data Extraction from XML WebPages" Published in IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 1, March, 2013.

[6] P.V.Praveen Sundar "Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural - Semantic Entropy" Published in International Journal of Advanced Research in Computer Science and Software Engineering

[7] Bing Liu, Robert Grossman, and Yanhong Zhai. Mining data records in web pages. In KDD ‟03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 601–606, New York, NY, USA, 2003.ACM Press.

[8]Sagnik Ray Choudhary,Prasenjit Mitra,Andi Kirk,Silvia Szep,Donald Pellegrino,Sue Jones,C.Lee.Giles."Figure Metadata Extraction From Digital Document", *ICDAR 2013. 12th International Conference on*, volume 1, pages 135–139. IEEE, 2013.

[9] Aanashi Bhardwaj and Veenu Mangat, "A Novel Approach for Content Extraction from Web Pages", IEEE Trans. Knowledge and Data 978-14799-2291-8/14/$31.00.IEEE2014.

