International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 7, July - 2014

Recommendation on Efficient Data Extraction for Web Data

Priyanka Patil
Department of Computer Engineering
St. Francis Institute of Technology
Mumbai, India

Asso. Prof.Anuradha G.
Department of Computer Engineering
St. Francis Institute of Technology
Mumbai, India

Prof. A. K.Sen
St. Francis Institute of Technology
Mumbai, India

***Abstract*—A web database is an organized listing of web pages, which can be queried or updated through World Wide Web (WWW).Web databases generate Query Result Pages (QRPs) in accordance to queries posted by users. Many applications necessitate the automatic extraction of data from these query result pages. The result from query result pages is very important for many web applications, which cooperates with multiple web databases. Web extraction tool automatically extracts data from QRPs .The data extracted using web extraction tool is aligned in a structured format using Cosine-Similarity. The aligned data is used for recommendation and text mining purposes.**

Keywords— cosine similarity, Data extraction, Data record alignment, Recommendation.

I. INTRODUCTION

Web databases are also called as online database; there are two types of web, surface web and deep web. The surface web has specific URLs. Search engine result pages consist of those URLs. Whereas deep web pages are dynamically generated in response to the user's query. After receiving a user's query, a web database returns the relevant data values, either in the form of structured or semi structured, encoded in web pages. Many web applications need the data from multiple web databases. For these applications to further utilize the data embedded in HTML pages, automatic data extraction is necessary. Once data values are extracted and organized in a structured manner, such as tables,they can be compared and aggregated. Hence, accurate data extraction is vital for these applications. [1].

Deep Data:-

The Deep Web is also called as Hidden Web. It comprises all information that resides in autonomous databases behind web portals. Web pages in the Deep Web are dynamically-generated in response to a query through a web site's search form and often contain rich content. [2]

Data Extraction:-

Data extraction is the process of retrieving data from unstructured or poorly structured datasources for further data processing or data storage. The majority of data extraction comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.[5]

Web Data Extraction:-

Web data extraction is the process of retrieving unstructured data from web pages and importing it into a structured data system like a database. Process of extracting data from Web pages is also referred as Web Scraping or Web Data Mining.[5]

Common Problem with Web Data Extraction:-

1. Incapable of processing with zero query results they require at least two records in a query result page.

2. Vulnerable to optional and disjunctive attributes It causes data alignment Problem.

3. Incapable of processing nesting data structures many methods can only process a flat data structure and fail for a nested data structure.

The aim of this work is to develop a system which provides efficient way to perform data extraction and data alignment. Data extraction is implemented using web extraction tool and for Data alignment string cosine similarity algorithm is implemented. Upon the aligned or structured data recommendation is given to the user.

The rest of this paper will be arranged as below. In section II will give the detailed information of web data extraction and alignment. In section III, the proposed work and overview of entire system. In section IV, result of data extraction is given and finally conclusion in section V.

II. REVIEW OF LITERATURE

Web database extraction is gaining popularity among the Database and Information Extraction research areas in recent years due to the volume and quality of deep web data. As the returned data for a query are embedded in HTML pages, the research has focused on how to extract this data. Earlier work focused on wrapper induction methods, which require human assistance to build a wrapper.[3]

The WWW is large repository of information on growing demand. It is very hard to query unstructured data. This is due to abundance of pages in unstructured data which is

generated dynamically from database.[1] Extraction of structured data is feasible for complex queries in the web page over the data which integrate the data present in different web-sites. [1] [2], In response to the queries, the database servers generate the information and deliver it directly to the user as Query Result Record (QRR). The generated information forms the hidden web (deep web or invisible web) and is usually enwrapped in Hypertext Markup Language (HTML) pages as data records. Due to the dynamic nature of the generated data records from the hidden web, current search engines (either general or commercial) are unable to index the HTML page accordingly. Thus, this type of web pages is termed deep web pages.

In[3],traditional data extraction from web pages uses the concept called wrappers" or "extractors". It extracts the contents of the web pages based on the knowledge of their formats which was developed manually in early time. A table extraction technique that works on web pages generated dynamically from a back-end database. System can automatically discover table structure by relevant pattern mining from web pages in an efficient way. Generate regular expression for the extraction process.[6]

III. PROPOSED WORK

The proposed work extracts QRR from QRPs in an automated fashion. The data extracted is in an unstructured format. The unstructured data is aligned using pairwise mechanism. The system is divided into following modules.

Data Extraction module: This module mainly focuses on extraction process using web extraction tool. i.e. Web Harvest Tool.

Data Alignment module: Once the data is extracted it is aligned with the help of pair wise alignment algorithm based on cosine similarity where accuracy of the result depends upon higher cosine similarity value.

Recommendation module: The final structured data i.e. QRR which in tabular format can be used to give recommendation to the user.

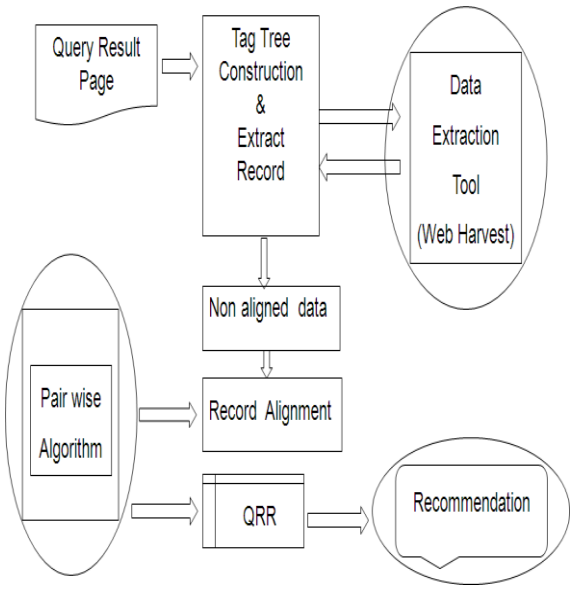

Figure 3.1.Proposed System

The proposed system flow is as shown in Fig 3.1 and System architecture of the proposed work is as shown in Fig. 3.2. The system is implemented for both static and dynamic websites.

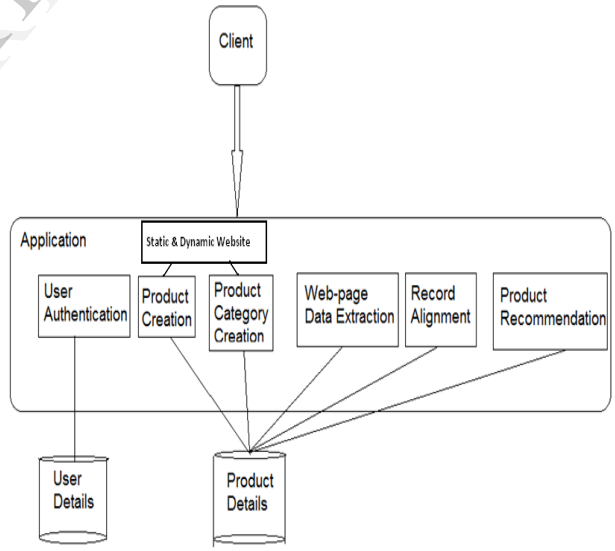

Fig 3.2. System Architecture

IV. RESULTS

The input to the proposed system is as shown in fig.3.1. yahoo shopping website. Fig 3.3 shows the snapshot of the website where the search string given is 'Camera'.

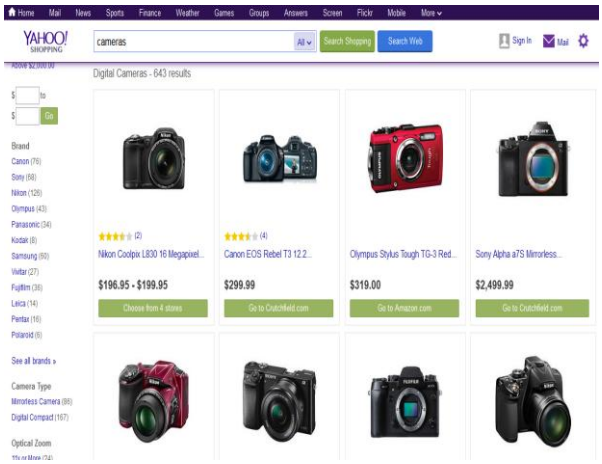

Fig 3.3 Search result for 'Camera

For Extraction of web data 'Web Harvest tool' is used. Web-Harvest is Open Source. Web Data Extraction tool written in Java language. It is used to collect Web pages and extract useful data from them. In order to do that, it clouts well established techniques and technologies for text/xml manipulation such as XSLT, Xpath and Regular Expressions. Fig 3.4 and Fig 3.5 shows snapshot of the 'Web Harvest tool'.

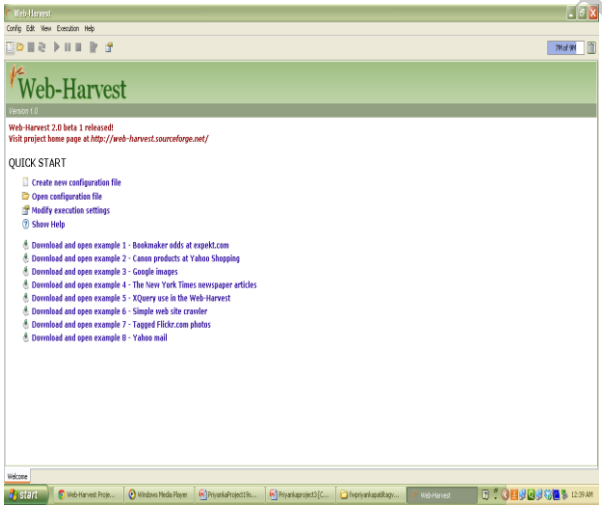

Fig 3.4 Web Harvest Tool

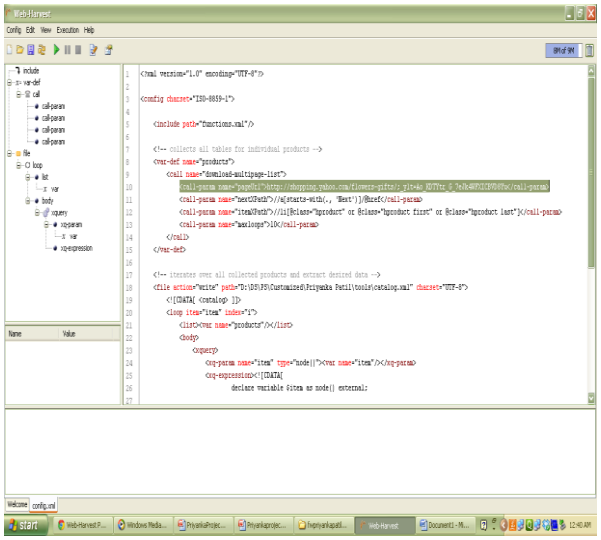

Fig 3.5. Website link for data extraction

Extracted data is in unstructured format as shown in Fig 3.6. This unstructured data is then aligned into structured format using pair wise algorithm that mainly focuses on tag and value similarity in HTML (Hypertext Markup Language) page. After alignment of the unstructured data it is transformed into structured data which is stored into a database in the form of rows and columns. Finally this tabular data can be used to give proactive recommendation to the users.

Attribute1	Attribute2	Attribute3	Attribute4
Canon EOS 7D 18 Megapixel...	$1,349.99 - $1,799.99		Choose from 2 stores
Canon EOS 70D (3) Black Digital SLR...			$1,149.99 - $1,349.99
Canon EOS 70D Black SLR Digital...	$1,199.00 - $1,549.99		Choose from 4 stores
Canon EOS 6D Black SLR Digital...	$1,399.00 - $1,899.99		Choose from 4 stores
Canon EOS-1D X Black SLR Digital...	$5,199.00 - $6,799.99		Choose from 2 stores
Canon EOS Rebel T3i 18 Megapixel...	$379.00 - $499.99		Choose from 2 stores
Canon EOS (3) Rebel T3i Black SLR...			$479.00 - $599.99

Fig 3.6 Unstructured data after extraction

V. CONCLUSION

From the query result pages the relevant data is extracted using the web harvest tool. The non-aligned data is to be aligned further using pair wise algorithm which can be used for recommendation purposes.

ACKNOWLEDGMENT

I would like to express my deep gratitude to Associate Professor Anuradha G., my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. I would also like to extend my thanks to the authors. Finally, I wish to thank my parents for their support and encouragement throughout my study.

REFERENCES

[1] Weifeng Su, Jiying Wang, Frederick H. Lochovsky, Member, IEEE Computer Society, and Yi Liu "Combining Tag and Value Similarity for Data Extraction and Alignment" *IEEE Transactions on knowledge and data engineering, vol. 24, no. 7, july 2012.*

[2] Jer Lang Hong "Data Extraction for Deep Web Using WordNet" *IEEE Transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 41, no. 6, november 2011*

[3] Mohammad Shafkat Amin Hasan Jamil, "FastWrap: An Efficient Wrapper for Tabular Data Extraction from the Web," *IEEE IRI* 2009, July 10-12, 2009, Las Vegas, Nevada, USA

[4] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," White Paper, BrightPlanet Corporation, http://www.brightplanet. com/resources/details/deepweb.html, 2001.

[5] Web Data Extraction. [Online] . Available: http://www.automationanywhere.com/solutions/webDataExt

[6] K.C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured Databases on the Web: Observations and Implications," *SIGMOD Record*, vol. 33, no. 3, pp. 61-70, 2004.

[7] P.V.Praveen Sundar Research Scholar"Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural - Semantic Entropy "*International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 4, April 2013*