International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 17 Issue 2 – SEPTEMBER 2015.

DATA FORMATTING AND ALIGNMENT USING PAIRWISE ALIGNMENT AND RECOMMENDATION SYSTEM

Shilpa Pote[1#] and Dr. D.R. Ingle [*2]

[#] *Student, Computer Engineering, Bharati Vidyapeeth COE, Maharashtra, India*
[*] *Dr. D.R.Ingle, Hod of Computer department, Bharati Vidyapeeth COE, Maharashtra, 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.**

Index Terms— *Data Extraction , Query result record .*

I. INTRODUCTION

Internet and the Web have revolutionized access to information. Today, one finds primarily on the Web, HTML (the standard for the Web) but also documents in pdf, doc, plain text as well as images, music and videos. The public Web is composed of billions of pages on millions of servers. It is a fantastic means of sharing information. It is very simple to use for humans. On the negative side, it is very inappropriate for accesses by software applications. This motivated the introduction of a semi structured data model, namely XML, which is well suited both for humans and machines .

1.1 Overview of web database

A web database is a database that can be queried and/ or updated through the World Wide Web (WWW).Web databases are also called as online databases. As web technologies are evolving, the WWW turned out to be the preferred medium for many applications, for example, e-commerce and digital libraries. These libraries use information that is stored in huge databases and can only be retrieved by issuing direct queries. A web database is a system for storing information that can then be accessed via a website. For example, an online community may have a database that stores the username, password, and other details of all its members. The most commonly used database system for the internet is MySQL due to its integration with PHP — one of the most widely used server side programming languages. The numbers of database-driven web sites are increasing exponentially. The dynamically created web pages by these sites are hard to be reached by traditional search engines. Traditional search engines are used to crawl and index static HTML pages. However, traditional search engines cannot send queries to web databases. The hidden information inside the web database sources is called the "deep web "in contrast to the"surface web "that is easily accessed by traditional search engines.

1.2 Surface Web

Surface web is that portion of the World Wide Web that is index-able by conventional search engines. Example of Surface Web pages may include Google, Facebook, YouTube, The New York Times, and other shopping websites. Surface Web is made up of static and fixed pages, and Static pages do not depends on a database for their content They reside on a server waiting to be retrieved, and are basically html files whose content never changes. Any changes are made directly to the html code and the new version of the page is uploaded to the server.

1.3 Deep Web

The Deep Web is also called as Hidden Web. 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.The Deep Web comprises all information that resides in autonomous databases behind portals and information providers' web front-ends. 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. A recent study has estimated the size of the Deep Web to be more than 500 billion pages, whereas the size of the "crawlable" web is only 1% of the Deep Web (i.e., less than 5 billion pages). Even those web sites with some static links that are "crawlable" by a search engine often have much more information available only through a query interface. Unlocking this vast deep web content presents a major research challenge.

1.4 Data Extraction

Data extraction is the process of retrieving data from unstructured or poorly structured data sources 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] After receiving a user's

query, a web database which is semi structured database returns the relevant data values, in structured format. 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.

1.5 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.

Common Problem with Web Data Extraction:-

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

- Vulnerable to optional and disjunctive attributes. It is true for those attributes which are not connected with each other or inconsistence in nature, such kind of attributes can cause data alignment problem.

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

II. LITERATURE SURVEY

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 from a query are embedded into HTML pages, the current research areas are focused on the extraction of this data. Earlier work focused on wrapper induction methods, which require human assistance to build a wrapper. The WWW is large repository of information on which grows in accordance with growing demands. This repository of unstructured data is very hard to query. This is due to abundance of pages in unstructured data which is generated dynamically from database. Extraction of structured data is feasible for complex queries in the web page where the data is integrated from different websites. 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 wrapped in Hypertext Markup Language (HTML) pages as data records.

III. ARCHITECTURE

In Proposed solution 3-tier architecture is used. Three-tier architecture is typically composed of a presentation tier, a business or data access tier, and a data tier.

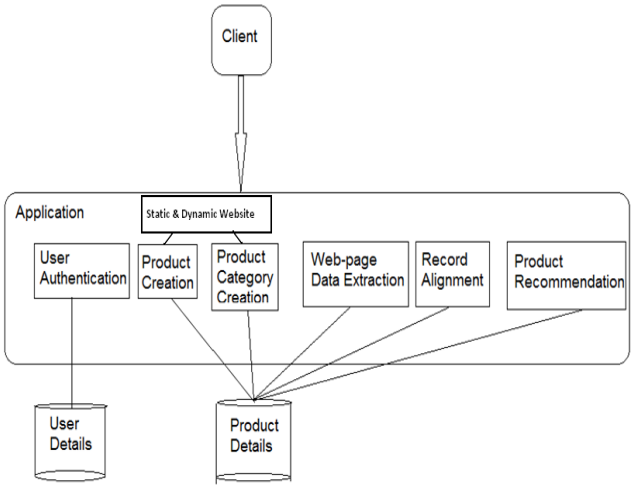

Fig. System Architecture

System architecture of the proposed work is as shown in Figure as shown in the figure The system is consists of different phases those are User authentication, product creation, Product Category Creation. They will save user details and product details respectively. Where, product and it's category creation is implemented for both static and dynamic websites.

Presentation tier

This is the top most level of the application. The presentation tier displays information related to such services as browsing merchandise, purchasing and shopping cart contents. It communicates with other tiers by outputting results to the browser/client tier and all other tiers in the network. (In simple terms it is a layer which users can access directly such as a web page, or an operating systems GUI)

Application tier (business logic, logic tier, data access tier, or middle tier)

The logical tier is pulled out from the presentation tier and, as its own layer; it controls an application's functionality by performing detailed processing.

Data tier

This tier consists of database servers. Here information is stored and retrieved. This tier keeps data neutral and independent from application servers or business logic. Giving data its own tier also improves scalability and performance.

IV. PROPOSE SYSTEM METHODOLOGY

4.1 Data Extraction

In the proposed system data extraction is implemented using web harvest tool which makes data extraction process efficient and fast. The steps involved in Query result record (QRR) are as follows:

4.1.1 QRR Extraction

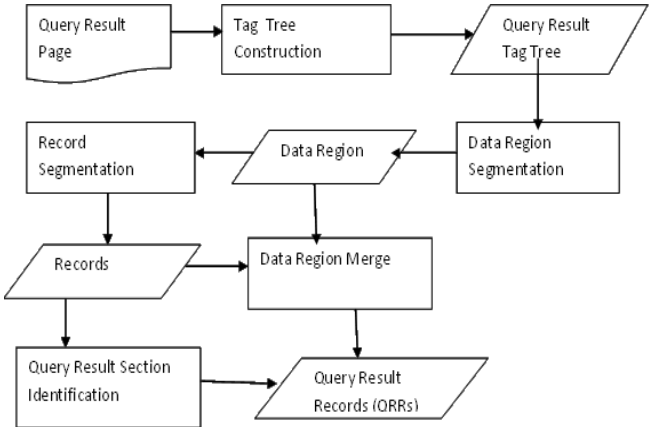

Fig. QRR extraction

Fig shows the framework for QRR extraction. Given a query result page, the Tag Tree Construction module first constructs a tag tree for the page rooted in the <HTML>tag. Each node represents a tag in the HTML page and its children are tags enclosed inside it. Each internal node n of the tag tree has a tag string tsn, which includes the tags of n and all tags of n's descendants, and a tag path tpn, which includes the tags from the root to n. Next, the Data Region Identification module identifies all possible data regions, which usually contain dynamically generate data, top down starting from the root node. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions. Given the segmented data records, the Data Region Merge module merges the data regions containing similar records. Finally, the Query Result Section Identification module selects one of the merged data regions as the one that contains the QRRs.

4.1.2 Web harvest tool

At present, existing web content is mainly formatted in unstructured HTML, even though flexible markup languages. e.g. XML, XHTML, are attracting a lot attention recently. Moreover, HTML is mainly presentation-oriented and is not really suited for database applications, while XML separates data structure from layout and provides more suitable data representation. Imagine a set of XML documents or structured tables can be regarded as a database and can be directly processed by a database application. But, we also know querying relevant data from unstructured HTML content spent a huge amount of time and cost. That is why we need a tool, in order to implement web data extraction

4.2 Data Alignment

Data alignment is implemented using pairwise algorithm. Pairwise algorithm used cosine similarity measures. They are totally depended on the similarity measure between two vectors. Here, two different data records are compared for their similarity. For those similarity and dissimilarity measures are as follows:

Similarity Measures:

Many data mining and analytics tasks involve the comparison of objects and determining in terms of their similarities (or dissimilarities)

Many of today's real-world applications rely on the computation similarities or distances among objects.

- Recommender systems
- Document categorization
- Information retrieval

Similarity and Dissimilarity:

Similarity

- Numerical measure of how alike two data objects are
- Value is higher when objects are more alike
- Often falls in the range [0,1]

Dissimilarity

- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0

4.3 Recommendation Module

In recent years we see the continuing growth of the Internet. Not only is the number of internet users and websites increasing, but also the amount of information on the individual websites. Many websites are concerned with presenting their often very semantically versatile information in a concise and efficient way. This is especially true for large E-Commerce websites with large amount of product information.

A frequently used technique to improve the presentation of data and navigation in these data is web recommendations. Web recommendations are hyperlinks, often augmented with short descriptive text and/or picture, which are shown on the website in addition to the usual content in order to lead users to potentially interesting information.

The motivation for the use of web recommendations comes from both internet users and website owners. Internet users want to see interesting information; the website owners want their information to reach users quickly and to the full extent. Owners of commercial websites also employ web recommendations in order to sell additional products or services to the users and thus increase the sales turnover of their websites.

V. CONCLUSION

The proposed approach has many modules. The first and the most important step is Data Extraction on web data. Web harvest tool, wampserver and ApacheTomcatServer have been used for implementing the data extraction phase of the proposed approach. The proposed system is gives faster and easy retrieval of data from the webpage using web extraction tool. This system extracts the result using web harvest tool which the existing system does not do.

The second step is Data alignment which makes use of cosine similarity mechanism. After data alignment the data records are stored in a tabular format. This approach adds efficiency to data extraction and alignment. It gives better

query performance. By using Recommendation module on the resultant records will provide personalized recommendation to the user on basis of their interest for particular product. Proactive analysis helps to give recommendation to the user.

REFERENCES

[1] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.
[2] R. Baeza-Yates, "Algorithms for String Matching: A Survey," ACM SIGIR Forum, vol. 23, nos. 3/4, pp. 34-58, 1989.
[3] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual Web Information Extraction with Lixto," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 119-128, 2001.
[4] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," White Paper, BrightPlanet Corporation, http://www.brightplanet. com/ resources/details/deepweb.html, 2001.
[5] P. Bonizzoni and G.D. Vedova, "The Complexity of Multiple Sequence Alignment with SP-Score that Is a Metric," Theoretical Computer Science, vol. 259, nos. 1/2, pp. 63-79, 2001.
[6] D. Buttler, L. Liu, and C. Pu, "A Fully Automated Object Extraction System for the World Wide Web," Proc. 21st Int'l Conf. Distributed Computing Systems, pp. 361-370, 2001.
[7] 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.
[8] C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proc. 10th World Wide Web Conf., pp. 681-688, 2001.
[9] L. Chen, H.M. Jamil, and N. Wang, "Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification," SIGMOD Record, vol. 33, no. 2, pp. 58-64, 2004.
[10] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.
[11] W. Cohen and L. Jensen, "A Structured Wrapper Induction System for Extracting Information from Semi-Structured Documents," Proc. IJCAI Workshop Adaptive Text Extraction and Mining, 2001.