Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 5, April-2019, ISSN 2230-7540

Web Scraping Based on Tag and Value Similarities

Vijay R. Thombare[1]* Prof. Shailesh Patil[2]

[1] Department of Computer Engineering, PG Student, Smt. Kashibai Navale College of Engineering Vadgaon Bk., Pune, India

[2] Department of Computer Engineering, Faculty, Smt. Kashibai Navale College of Engineering Vadgaon Bk., Pune, India

Abstract – Web Scraping is a technique for extracting huge amounts of data available on internet websites. The text data available on websites is generally not available to download directly and can't be used for some other application. It's only accessed by using a web browser via their HTML query interface. A web page also contains irrelevant data, such as advertisements, comments, GIF and other links. We are presenting a technique to automatically extract result records from the dynamically generated result page returned by search engine. This paper present an efficient extraction and alignment procedure called EXCTVS which considers both tag and value likeness. It extracts data from query result pages by first recognizing and then segmenting the Query Result Records (QRRs) based on its tag and value considering tags similarities. Once extraction is completed, it aligns the segmented QRRs into a table. This paper put data values with identical attribute into the identical column. This paper suggests a new method to handle the case where the QRRs are not contiguous in web pages, which may be due to the occurrence of auxiliary data such as a comments, recommendation or promotion. This paper is considering the nested structures of web pages while processing the QRR. This paper uses the record alignment algorithm that aligns the attributes in a record, first it do by pair wise and then holistically, by combining the tag and tag values similarity information.

Keywords: Automatic Wrapper Generation, Data Extraction, Data Record Alignment, Information Integration.

----------X----------

1. INTRODUCTION

Web is the largest repository of open data and this data has been growing at exponential rates since the inception of internet. Data record mining in web pages is very important because they typically present their host pages basic information, such as product lists and services details. Extracting these structured data objects enables to integrate data from multiple web pages available over internet to provide value-added services, e.g. shopping sites, search engines. These structured data objects are important type of information on the Web (HTML pages). These data objects which holds data in a web page are nothing but the records from underlying databases and displayed in web pages with some fixed templates and style. In this paper, this paper is going to call them as data records. The data available on websites is generally not available to download easily and can only be accessed by using a web browser. Through these web pages the proposed method indirectly accessing it's underneath database without any access permissions as this all data is available on web itself. This data can be used for the comparison or some analysis purpose.

Web data is of great use to Ecommerce portals, media companies, research firms, data scientists, government and can even help the healthcare industry with on-going research and making predictions on the spread of diseases.

Consider the data available on classifieds sites, real estate portals, social networks, retail sites, and online shopping websites etc. being easily available in a structured format, ready to be analysed for more purpose. Most of these sites don't provide the functionality to save their data to a local or cloud storage for future use. Also these most of sites don't allow direct data comparison with some other site data. Some sites provide APIs, but they typically come with restrictions and aren't reliable enough. Although it's technically possible to copy and paste data from a website to your local storage, this is inconvenient and out of question when it comes to practical use cases for businesses. Web scraping helps you do this in an

automated fashion and do this efficiently and with accurately. A web scraping setup interacts with web page in a way similar to a web browser does, but instead of displaying it on a screen, it saves the required data to a storage system.

Data mining concept is now getting famous day by day as internet uses are increasing. Data mining is process of extracting hidden and valuable data and information from large data bases. It involves methods and algorithms to extract knowledge and data from different data repositories such as transaction databases, data warehouses text files, and web database etc., as these are actual sources of data which is available on web pages. From the beginning of year 1990, World Wide Web has grown exponentially in its volume size. In (Chang, et. al., 2004), it is estimated that it contains approximately 50 billion publicly remote accessible index able web documents distributed all over the world on thousands of web servers. It is a very roasting job to search information from such a big big collection of web documents on World Wide Web as the web pages and documents are not organized as books on shelves in a library, nor are web pages completely stored at one central location. It is not guaranteed that users will be able to extract the information even after they know where to look for information at its URLs as web is constantly changing along with its data contains. Therefore, there was a need to develop some information extraction tools to search the required information from WWW. Web information extraction tools are mainly divided into three categories as Web directories, Meta search engines, Search engines.

2. RELATED WORK

Recent studies say World Wide Web (www) serves a huge, widely distributed, global information providing services which are increasing rapidly. Much more information is presented in the form of a web record which exists in both detail and list pages. Due to the increase of online web databases, it is good to have useful required information which will be formatted before presenting to the users and this is one of the web information extractions (WIE) tasks. In this paper have studied many different methods to do web data extraction and scraping methods such as Viper method (Simon and Lausen, 2005), wrapper induction (Zhao, et. al., 2005), Road-runner (Valter Crescenzi, et. al., 2002) and automatic object extraction system (Buttler, et. al., 2001), however each of them having limitations (Crescenzi et. al., 2001), (Liu, et. al., 2003), (Zhao, et. al., 2005).

These typical information extraction tasks focus on data regions and data records. It implies that as the complexity of typical web documents increases day by day with new technologies, information extractors have to analyze more and more irrelevant regions to retrieve the relevant information only and exclude the unwanted contains, with consideration of both efficiency and effectiveness. This has motivated a number of authors to work on region extractors as a means to relieve information extractors from the burden.

There are two main approaches for data extraction. The first one is the wrapper induction (Liu, et. al., 2003). Wrapper uses supervised learning to learn data extraction rules from a set of manually labeled positive and negative examples. It include manual labeling of data is which is labor intensive and time consuming. Also, for different sites or even pages in the same site, the manual labeling process needs to be repeated because they all mostly follow different templates and patterns. Example wrapper induction systems include WIEN (Kushmerick, 2002), WL2 (Cohen, et. al., 2002) are inefficient as it need manual efforts for labeling of data and time consuming. Our technique requires no human labeling. (Liu, et. al., 2003) are unable to process the nested structure of HTML tags, while our system process nested structures also.

The other method is automatic object extraction from web pages called Omini (Buttler, et. al., 2001). It performs extraction in two stages by parsing web page into tree structure. In first it use algorithm and locate smallest sub tree that contains all the object of interest. Second step finds the correct object separator tags that can separate object efficiently using algorithm. This method is based on a set of heuristic rules, like highest count tags, standard deviation, repeating-tags and ontology-matching. Also (Buttler, et. al., 2001) Proposes a few more heuristics to perform the task without using domain ontology. But in the end this method shown that this one produce poor results. In addition, these methods were unable to extract data from noncontiguous data records. Our new method does not use tag strings for alignment but trees. This gives an edge over all the other methods because it exploits nested tree structures to perform much more accurate data extraction and also gives set of heuristics to find individual product information, e.g. price and others. In (Weifeng, et. al., 2012), (Crescenzi et. al., 2001), two more techniques are proposed. However, these old techniques needed multiple pages which are assumed to be given, that contain similar data records from the same site to find patterns and grammars from the pages to extract data records. Current existing methods assume that the availability of multiple pages containing similar common data records is a big limitation. Our method works on each single page. In (Weifeng, et. al., 2012) they proposed a system which works for the non-contiguous date regions and processes the nested structures also. But this method requires at least two QRR in the query result pages (Gitanjali Shirsath et. al., 2017), (Paul, 2015) proposed system to extract data for unsupervised web document. But this system uses regular expression for identifying the extraction dataset. This system based on the regular

expression, so its work for the fixed and similar web pages only. These regular expressions are not able to handle the nested HTML page tags. It will works only on simple HTML pages not with the Nested XHTML pages.

3. PROPOSED SYSTEM

A. Architecture

We propose a novel method for data scraping from the structured and unstructured, contiguous and noncontiguous web pages. Proposed a new technique based on combined tag and value similarity for the extraction of QRRs from a query result pages, which works in two stages as record extraction and its alignment. This system mainly focuses on the problem of automatic date extracting from web that are encoded in the query result pages generated by web databases. In general, a query result page containing the actual data along with the other information which should not be considered while extraction, such as navigational panels, animations, advertisements, comments, information about hosting sites. The goal of web database data extraction is to remove all irrelevant and unnecessary information from the query result page and extract only useful information in the query result records from the page. This method aligning the extracted QRRs into a table such that the data values which belong to the same attribute are placed into the same table column.

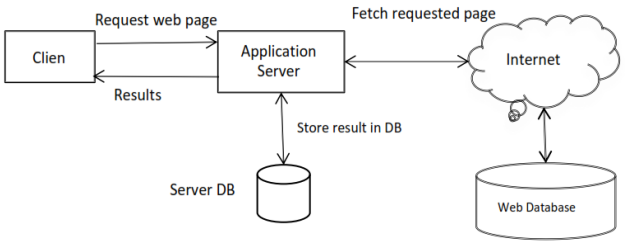

Fig 1- Architecture Diagram

Fig. 1 shows the proposed system architecture for the web scraping method based on tag and value similarity. Here in this architecture, three main actors are Client, Application server and Web database. Application server takes the input from the client i.e. web page request. Application server will fetch that page from the web database using the page URL submitted by the client. After fetching the whole page, the application server will extract the data from page. The extraction is done using the extraction algorithm. The result of the extraction is in tabular form which is stored in the database B. Proposed Algorithm In Proposed work here, it present a novel web data scraping method, to automatically extract QRRs from a query result page. It employs two steps for this task. In first step is to identify and segment the QRRs. This paper is allowing the QRRs in a data region to be non-contiguous to improve on existing techniques. The second step aligns the data values among the QRRs. This paper proposed a novel alignment method is in which the alignment is performed in three steps: pair wise alignment, holistic alignment, and nested structure processing.

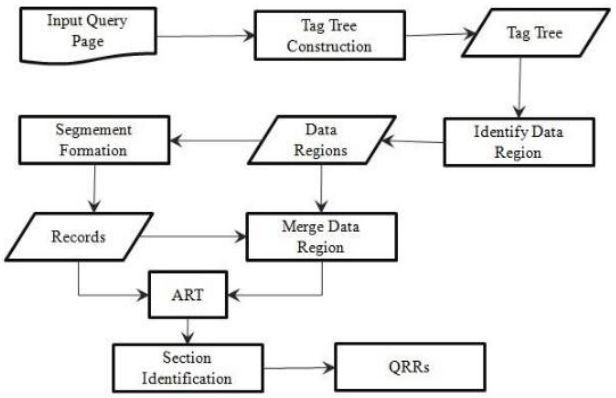

Fig. 2 – Data Extraction framework

We introduce a new technique in this Web Scraping for the extraction of QRRs from a query result page. Record extraction identifies the QRRs in a query result page which involve the following sub steps:

1) Data region identification, buffering, semantic extraction and the segmentation step.

2) Record Alignment where the data values for the same attribute are aligned and put in to the same column of the table.

Comparing with the existing CTVS (Zhai and Liu, 2006) technique, this improves the data extraction accuracy in two ways:

1) Optional labelling is the technique by which the problem of elimination of optional attribute that appears as the start node in a data region, as auxiliary information is eliminated. This is incorporated in the record extraction step.

2) Dynamic tagging is the other improvement. The existing system uses static tagging which results in less accurate results. The dynamic tagging uses the semantic data extraction concept. In the static tagging only the attributes and values recorded in prior can be used.

3) The existing system uses the datasets as result pages which are previously stored on local machine. It should have at least two QRRs for the extraction purpose. This proposed method overcomes this.

Web Scraping Algorithm:
Input: Query Result Record, R.
Output: Extracted Data, E.
Step 1. Input Query.
Step 2. From the available links find the keywords.
Step 3. Store the information to a database.
Step 4. Perform structure analysis.
Step 5. Extract tags from the link.
Step 6. Store them to a temporary file.
Step 7. Match the attributes and identify the data regions.
Step 8. Segment the records, Temp Containing optional data QRR Actual records.
Step 9. Merge QRRs.
Step 10 If the result not found then go for semantic extraction.
Step 11 Repeat step 5.
Step 12 Final Result section is identified

Table 1 – Web Scraping Algorithm

We are using ART Algorithm for alignment purpose in this system. ART in particular was designed for resolving the stability-plasticity problem. Here it refers to conflict in the ability to maintain old learned information while still being adaptive to learn the new information. An algorithm is defined as plastic if it can adapt to new information. Additionally this algorithm is stable if it can retain previously learned knowledge. Our goal is to create an algorithm that can also retain previously learned knowledge while at the same time integrating newly discovered knowledge. In this way, the algorithm is both staple and plastic.

4. RESULTS AND DISCUSSION

Web page data extraction involves extraction of relevant data from HTML web pages. Thus this system here is taking the unstructured data from HTML tags and scraping this data into structured database format based on tags and its value similarities. Fig.1 shows the general view of the process this system is following here for the scraping purpose. The data scraping algorithm, proposed here is as given in table 1, in this system algorithm is implemented using the .net framework.

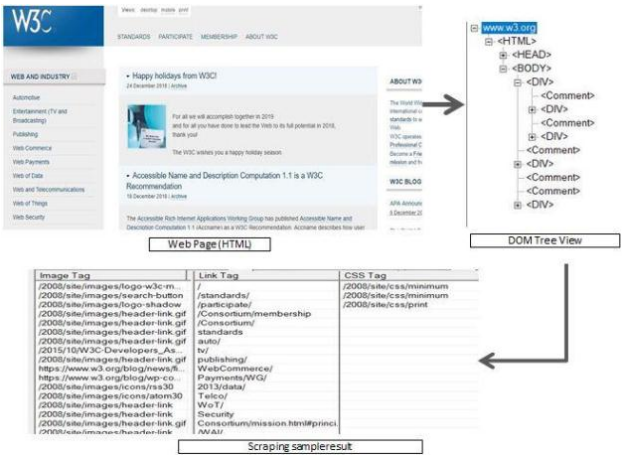

Fig. 3 – Data scraping system result

Fig.4 shows us how system is working here. Input to this system is collection of web pages with sample template as in sample system had taken w3.org homepage. It is parsing the HTML input page for multiple tags in it and forming a DOM tree view for page. Based on DOM tree tag values this system will be extracting the specific tag data values, here in this system had extracted image, link and css tag values. These same result set will be stored in database for further use. This build on the hypothesis of extraction of required data from web page considering predefined HTML tags for extraction. Using algorithm in Table 1, it first identifies the data region in given web page and then it extract region data into its corresponding QRR. Once all region data is extracted this will marge these QRR data into single one and result is stored in database as a final result of the system.

5. CONCLUSION

In this Paper tried to make such an algorithm for web scraping, which automatically extract QRRs from a query result page. It works in two steps for performing the resultant task. The first step identifies and segments the QRRs. In which this system have improved on existing techniques by allowing the QRRs in a data region to be non-contiguous. The second step aligns the data values among the QRRs. A novel alignment method for web scraping is proposed in which the alignment is performed in three consecutive steps: pair wise alignment, holistic alignment, and nested structure processing. This system will be using ART algorithm because ART algorithm is ability to create a new cluster if the underlying data warrants.

REFERENCES

1. Gitanjali Shirsath et. al. (2017). International Journal of Computer Science and Mobile Computing, “INFORMATION EXTRACTION FROM WEB PAGES USING PATTERN MATCHING” ,Vol.6 Issue.3, March- 2017, pg. 14-19.

2. Roselin Paul (2015). “Automated Data Extraction for Unsupervised Web Documents", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015, pg. 10613-10618.

3. Weifeng Su, Jiying Wang, Frederick H. Lochovsky and Yi Liu (2012). “Combining Tag and Value Similarity for Data Extraction and Alignment”, VOL. 24, NO. 7, JULY 2012.

4. K. Simon and G. Lausen (2005). ViPER: Augmenting Automatic Information Extraction with Visual Perceptions", Proc. 14th ACM Intl Conf. Information and Knowledge Management, pp. 381-388.

5. H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu (2005). "Fully Automatic Wrapper Generation for Search Engines", Proc. 14th World Wide Web Conf., pp. 66-75.

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

7. Liu B., Grossman R. and Zhai Y. (2003). "Mining data records from Web pages." KDD-03.

8. Valter Crescenzi, Giansalvatore Mecca and Paolo Merialdo (2002). "Automatic Web Information Extraction in the road Runner System", pp. 264–277, Springer-Verlag Berlin Heidelberg.

9. Cohen W., Hurst M., and Jensen L. (2002). "A flexible learning system for wrapping tables and lists in HTML documents". WWW-2002.

10. Y. Zhai and B. Liu (2006). "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. D. Buttler, L. Liu, and C. Pu (2001). "A Fully Automated Object Extraction System for the World Wide Web", Proc. 21st Intl Conf. Distributed Computing Systems, pp. 361-370.

12. Alberto H. F. Laende, R. Berthier A. Ribeiro Neto, Altigran S. da Silva, Juliana S. Teixeira. "A Brief Survey of Web Data Extraction Tools" Department of Computer Science, Federal University of Minas Gerais, 31270901,Belo Horizonte MG Brazil.

13. Crescenzi V., Mecca G. and Merialdo P. (2001). "Roadrunner: Towards automatic data extraction from large web sites". Proc. 27th Intl Conf.Very Large Data Bases, pp. 109-118.

14. Kushmerick N. (2002). "Wrapper induction: efficiency and expressiveness. Artificial Intelligence",118: pp. 15-68.

15. Hsu C.N. and Dung M.T. (1998). "Generating finite-state transducers for semi-structured data extraction from the Web. Information Systems". 23(8): pp. 521-538.

16. Hammer J., Garcia-Molina H., Cho J., Aranha R., and Crespo A. (1997). "Extracting semi-structured information from the Web. Workshop on Manag. of Semi-structured Data."

Corresponding Author

Vijay R. Thombare*

Department of Computer Engineering, PG Student, Smt. Kashibai Navale College of Engineering Vadgaon Bk., Pune, India

vjthombare0205@gmail.com