ISSN:0975-9646

A Survey on Data Extraction of Web Pages Using Tag Tree Structure

Vivek D. Mohod, Mrs. J. V. Megha
Department of Information Technology
Shri Guru Gobind Singhji
Institute of Engineering & Technology
Nanded, 431606

***Abstract*— Internet contains large amount of data which user want to retrieve with the help of search input query. But the result return from the web has multiple dynamic output records. Hence, there is need of flexible information extraction system to convert web pages into machine process able structure which is essential for much application. This, essential information need to be extracted & annotated automatically which is challenge in data mining. In this paper, we survey on different HTML structure based technique to scrap data from web pages.**

***Keywords*— Data records, data extraction, HTML structure, unstructured web pages.**

I. INTRODUCTION

Search engine provide some help for user in looking information from internet. But the web pages returned in search forms are not properly indexed. Thus, there are various technology and researches done which are focusing on data extraction from web data storage.

For example, if a user wants information about notebook, then such type of information only exists in back end database of various notebook vendors [3]. Each time user has to visit the web page of that respective vendor and collect the relevant information and compare them manually. To collect this information from different web sites, wrappers have been developed to extract data objects from web pages by some web information extraction systems. Such information extraction systems need human involvement which does not provide efficiency in assigning meaningful attribute to the extracted data.

So, there is need to solve this problem by automatically extracting data from given web site and assign meaningful label to extracted data objects from a web. Extracting structured data provides us facility to combine data from multiple webpages. Researchers have developed different automatic wrappers Eg. Stalker[8], Softmealy[9] etc. Existing approaches depends on the template of the web pages which require structure finding of the template in that webpage. Previous work on information extraction systems requires lot of human efforts for annotating data units. Our approach focus

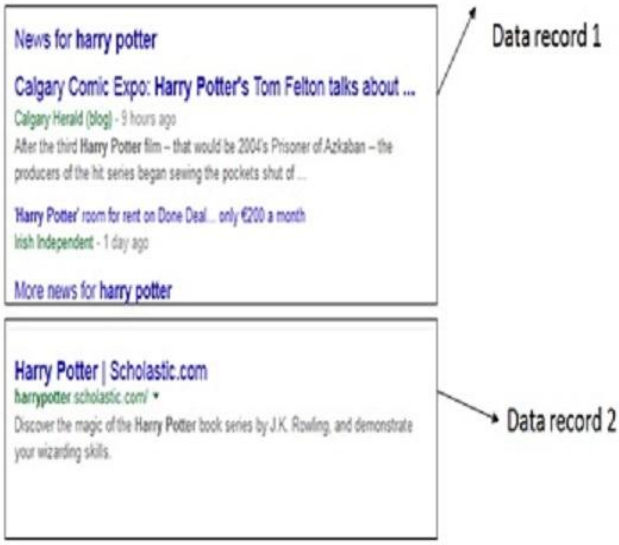

Fig. 1. Sample webpage containing multiple data sections

on tag tree structure of webpage to identify repeated pattern, we divide the page into sections which contains all data record by analysing HTML tag tree then the grammar is generated using tag tree to extract data. So, this reduces human involvement and increase accuracy.

II. LITERATURE SURVEY

Information extraction provides services to user which retrieves the information by firing query on internet. But this approach is not effective to produce accurate result because human involvement and poor quality of data extraction output. The important task for information extraction depends on its input and extracted target. Thus, input can be free text which is written in natural language or semi-structured document which is arranged in tables or enumerated lists [6]. For extracting information from attributes, it may have zero or multiple records.

There are two problems while extracting information mainly first: to categorize unstructured view of data. Second: to categorize structure and semi-structure view of data. Many systems need the program or the set of rules for extracting structured data from web using wrapper. This wrapper requires human involvement which results in poor

scalability and was not suitable for application. And if, the format of web page changes then it becomes difficult for wrapper to maintain which is expensive. We require semantic grouping of extracted information to make web data machine process able. The semantic grouping means data with similar meaning which form with same concept.
Using data tree algorithm based on regular expression DELA first uses HTML tags to align data units by filing them into table. Thus, we align data units and annotate the ones within same semantic group [2]. VIDE perform alignment and generates alignment wrapper using visual features. But this alignment was only at text level, not at data level. Based on various techniques 3 types of approaches have been analysed [4].In earlier phase, manual approach were designed to identify and extract desired data items by constructing wrapper. This was overcome by semiautomatic approach which uses sequence based and tree based techniques, but this requires manual efforts which are labour intensive time consuming [4]. So to reduce manual efforts and increases the efficiency, recent researches focuses on automatic approaches. Some automatic approaches perform only data record but not data item extraction. These approaches do not generate wrapper rather they perform extraction on web page directly.

III. FUNDAMENTAL

A. *Web Crawler*

Web crawler is tool of automatic collecting data or required information from web. It is software technique of extracting information from websites. Crawler process the HTML of web page, it restores the values from database of task specific concepts and assign those values as queries to form elements if the form labels and database elements match [3].

B. *Wrapper*

Wrapper is a program or set of rules which collects the pages generated by web crawler. It generates regular expression for HTML pages.

C. *Data Aligner*

Generated web page using wrapper, data objects are extracted from web pages by data aligner. By matching the wrapper with taken sequence it filters HTML tags and rearranges the data instances into manner, where rows represent data instances and columns represent attributes.

D. *Label Assigner*

The main task for label assigner is to assign label to the extracted data objects obtained in column of table by matching form labels. The basic idea is that the query word submitted through form elements which probably reappears in the corresponding fields of data objects, since web sites usually try their best to provide the most relevant data back to users [3].

IV. DATA EXTRACTION BASED ON TREE STRUCTURE TECHNIQUES

A. *IEPAD*:

IEPAD [11] extract the information from the similar web pages using extractor module. From the input web page repetitive pattern is discovered with the help of pattern viewer. This pattern viewer uses extraction rule for generating the results, extraction rule includes translator, PAT tree constructor pattern discovery and pattern validator. The GUI enables user to view the information extracted by extraction rule. From this GUI, user can select required information then extractor module can use this information for extracting data from other pages having similar structure.

B. *Record Boundary:*

Mostly web documents are defined in HTML structure that includes plain text and tags. We can define tags in web document as D and regions as R. Based on nested structure of web document, a tag tree is constructed to detect the region where the records of interest contains. This approach consist of following steps [12]

1. It develops extraction rule of model for domain of interest.
2. Using extraction rule database scheme is generated as well as rules for matching constants and keywords.
3. Record extractor is used to obtain data from web page that separates an unstructured web document in individual record-size. It then cleans it by removing markup language tags.
4. Recognizers are used for matching rules generated by parser to extract from cleaned individual unstructured documents from which we obtain desired output. The output is arranged in data record table.

C. *DELA:*

DELA [3] works on two steps for generating wrapper

1. Data-rich section extraction(DSE) algorithm and
2. Pattern extractor

DSE is designed to extract data-rich section from web page by comparing the DOM trees for two web pages of same web site and discarding the nodes with identical sub trees. Whereas, pattern extractor is used to discover continuously repeated pattern using suffix trees. The collection of web pages is given as input to wrapper generator. Wrapper generator produces regular expression based on HTML tag structure of the page. Wrapper generator consider each page as sequence of token composed of HTML tags. Repeated HTML tags are extracted and regular expression wrapper is derived from repeated substring. DELA also includes data aligner which also uses data extraction and attribute separation. Data extraction extracts data from web pages by using wrapper produced by wrapper generator. In attribute separation phase, several attributes are encoded into one string but then there must be special symbol in the string as the separator to visually support user to separate the attribute.

D. *ROADRUNNER:*

RoadRunner uses HTML pages for generating wrapper by visualising grammar for HTML code. The system uses the ACME matching technique to compare HTML pages of same class and generate a wrapper based on their similarities and differences [13]. RoadRunner starts with input page as its initial page as its initial template then it modifies its current template. Extraction process in RoadRunner is based on comparison of tag structure of sample pages.

E. *DEPTA:*

Web pages having two or more data records in a region, DEPTA (Data extraction based on partial tree alignment) [5] compares all adjacent substring with starting tags having same parent in HTML tag tree. The substring comparison is done with the help of tree edit distance where the similar substring are identified rather than matching exact string. The algorithm used in DEPTA first builds an HTML tag tree for the input page then it compares substring for all children under same parent where we find data region. Similar node are denoted by generalized nodes. At last data items from data records are extracted using partial tree alignment.

F. *Mining Data Records in Web Pages*:

This proposed technique has 3 step to extract data

1. Building HTML tag tree
2. Mining data regions and
3. Identify data records

After building HTML tag tree, data region that contains similar data records from web page is mined. Instead of extracting data records first, it extract generalized nodes in a page (similar nodes are denoted by generalized nodes). Then from each data region, actual data records are identified.

CONCLUSION

In this paper, we survey on different techniques of data extraction from web document to extract information. These technique are based on HTML structure, some technique identifies the data record without extracting data field, and some are based on visual information to extract data. Some techniques uses DOM tree to extract repeated pattern then this repeated pattern is used to extract data.

ACKNOWLEDGMENT

I would like to thanks my guide Mrs. J. V. Megha for her support, encouragement and valuable advices on this paper.

REFERENCES

[1] Yogesh W. Wanjari, Dipali B. Gaikwad, Vivek D. Mohod, Sachin N. Deshmukh, “Data Extraction and Annotation for Web Databases using Multiple Annotators Approach- A Review” International Journal of Computer Applications (0975 - 8887) Volume 88 - No.18, February 2014

[2] Y. Lu, H. He, H. Zhao, W. Meng, C.Yu, “Annotating Search Results from Web Databases”, IEEE Knowledge and Data Engg”., vol. 25, March-2013.

[3] J. Wang and F.H. Lochovsky, “Data Extraction and Label Assignment for Web Databases,” Proc. 12th Int’l Conf. World Wide Web (WWW), 2003.

[4] Wei Liu, Xiaofeng Meng and Weiyi Meng, “ViDE: A Vision-Based Approach for Deep Web Data Extraction”, IEEE transactions on knowledge and data engineering, vol. 22, no. 3, march 2010.

[5] Yanhong Zhai, Bing Liu, “Web Data Extraction Based on Partial Tree Alignment” WWW 2005, May 10-14, 2005, Chiba, Japan. ACM 1-59593-046-9/05/0005.

[6] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis and Khaled Shaalan, “A Survey of Web Information Extraction Systems” IEEE, TKDE-0475-1104.R3.

[7] Tai, K. The tree-to-tree correction problem. J. ACM, 26(3):422-433, 1979.

[8] I. Muslea, S. Minton, and C. Knoblock, A hierarchical approach to wrapper induction, In Proceedings of the 3rd International Conference on Autonomous Agents (Agents .99), Seattle, WA, 1999.

[9] C-N. Hsu and M-T. Dung, Generating finite-state transducers for semi-structured data extraction from the Web Information Systems, 23(8):pp. 521538, 1998.

[10] Vinayak B. Kadam , Ganesh K. Pakle “DEUDS: Data Extraction Using DOM Tree and Selectors” Vinayak B. Kadam et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1403-1410.

[11] C.-H. Chang, and S.-C. Lui, “IEPAD: Information extraction based on pattern discovery,” Proceedings of the Tenth International Conference on World Wide Web (WWW), Hong-Kong, pp. 223-231, 2001.

[12] D.W. Embley,Y.S.Jiang,Y.K.Hg,"Record-boundry Discovery in Web Documents".

[13] Crescenzi, V., Mecca, G. and Merialdo, P., RoadRunner: towards automatic data extraction from large Web sites. Proceedings of the 26th International Conference on Very Large Database Systems (VLDB), Rome, Italy, pp. 109-118, 2001.