www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 4 Issue 5 May 2015, Page No. 12040-12044*

# Web Data Extraction Using Partial Tree Alignment

***M.Mangala Nachar[1] B.Vinitha Subhasini[2]***
PG Student,Department of Computer Science and Engineering
Akshaya College of Engineering & Technology ,Kinathukadavu
Assistant Professor, Department of Computer Science and Engineering
Akshaya College of Engineering & Technology ,Kinathukadavu

***Abstract***--- *As a huge data source the internet contains a large number of worthless information, and the data of information is usually in the form of semi-structured in HTML web pages. This paper uses a new methodology to perform the task automatically. It consists of two steps, the foremost one is identifying individual data records in a page, and the next is aligning and extracting data items from the identified information data records. For the foremost process, A method based on visual information to segment of data records, which is more relevant to past methods. The next process, uses a novel partial alignment is a technique based on hierarchical parent child matching method. Partial alignment means we aligning only those data fields in a pair of data records that can be aligned (or matched) , and make none relevant information on the rest of the data fields. This approach does enables more reliable alignment in the multiple data records. The results are using a large number of Web pages from diverse domains show that the proposed two-step technique is able to segment information data history, align and retrieved data from the very well matched with relevant result. The parameters used are precision, recall and f-measure is used for evaluating the performance of the existing and proposed methods of web data extraction method. The process results prove that the proposed method is better than the existing method.*

*Keywords: Query Result Records, Parallel Automatic Deep Web Data Extraction algorithm method, web data extraction, Partial alignment technique*

## 1 Introduction

The explosive growth and popularity of the world-wide web has resulted in a huge amount of information sources in the Internet. However, due to heterogeneity and the lack of structure of Web information content, for this collection of information access has been limited to browsing and searching. For decision making many business applications have to depend on web in order to aggregate information from many web sites. By analyzing and discussing web data we can find latest market trends, cost details, product information etc. Web Data Extraction systems are a broad class of software applications targeting at extracting information from Web sources [1]. A Web Data Extraction system usually interacts with a Web source and extracts data stored in it: for instance, if the data source is an HTML Web page, the extracted information to consist of elements in the page as well as the full-text of the page by-self. Dispute, retrieved data might be Later-processed, converted in the most convenient structured format and stored for later use. Manual data extraction is takes more time and it reports in error prone.

Web Data Extraction systems find extensive use in a wide range of applications including the analysis processing , Business process and Competitive department, creep of Social Web platforms [2], Bio-Informatics and so on. The importance of Web Data Extraction systems depends on the fact that large (and steadily growing) amount of information is contiguously produced, exchange and consumed online: Web Data Extraction systems allow to efficiently collecting this information with limited human efforts. The availability and analysis of collecting data is an indefeasible requirement to understanding the complexity of social, scientific and economic phenomena which generate the information itself. For example, collecting digital traces produced by users of Social Web platforms like Face book, YouTube or Flicker is the key step to understand, model and predict human behavior [3].

Automatic data extraction plays an important role in processing results provided by search engines after submitting the query results by user. Wrapper is an automated tool which extracting Query Result Records (QRRs) from HTML pages returned by search engines. Automated extraction is easy with the sites having web service interfaces like Google and Amazon. But it's difficult for those support B2C applications which is not have web service interfaces. Normally Search engine results consists of query independent contents [4], query dependent contents , while some contents are affect by many queries but independent of content of specific query . As the web evolved web page creation process is changed from manual to more dynamic procedure

using complex templates. Many WebPages are created in advanced, but are generated dynamically by querying a database server and sending the results to a predefined page structure. Automatically data extraction is most important for many applications, and comparison shopping, that needs to co-operate with many web databases to collecting data from multiple sites and provide services.

Web Data Extraction is an important problem that has been studied by means of different scientific tools and in a broad range of applications. Some of the problems arises during the web extraction is Web Data Extraction techniques often require the help of human experts. Web Data Extraction techniques should be able to process large volumes of data in relatively short time. This requirement is particularly stringent in the field of Business and Competitive Intelligence because a company needs to perform timely analysis of market conditions.

The goal of this survey is to providing a structured and comprehensive overview of the research in Web Data Extraction as well as to provide an overview of most recent results in the literature.

## 2.Problem Solving Strategies

The goal of this survey is to providing a structured and comprehensive overview of the research in Web Data Extraction as well as to provide an overview of most recent results in the literature.
This paper proposed a two-step strategy to solve the problem.

1. Given a page, the method first segments the page to identify each data record without extracting its data items. We have improved our previous technique MDR for this purposed. Specifically, the new method also uses visual cues to find data records. Visual information helps the system in two ways:

(i) It enables the system to identifying gaps that separate data records, which help to segment data records correctly because the gap within a data record (if any) is typically smaller than that in between data records.
(ii) The proposed system identifies data records by analyzing HTML tags trees or DOM trees. A straightforward way to build a tags tree is to follow the nested tag structure in the HTML code. However, sophisticated analysis has to be incorporated to handle errors in the HTML code (e.g.,missing or ill-formatted tags). Whereas the visual or display information can be obtained after the HTML code
is rendered by a Web browser, it also contains information about the hierarchical structure of the tags. In this work, rather than analyzing the HTML code, visual information (i.e., the locations on the screen at which tags are rendered) is utilized to infer the structural relationship among tags and to construct a tag tree. This method leads to more robust tree construction due to the high error tolerance of the rendering engines of Web browsers (e.g., Internet Explorer). As long as the browser is able to rendering
the page correctly,and the tag tree can be built correctly.

2. A novel partial tree alignment method is proposed to align and to extract corresponding data items from the discovered data records and put the data items in a database table. Using tree alignments are natural because of the nested (or tree structured) organization of HTML code. This new method is very accurate as our experiments are show. Specifically, after all the data records have been identified, the sub-trees of each data record are re-arranged into a single tree as each data records may be contained in more than one sub trees in the original tags tree of the page, and every data records may not be continuously.

The tag trees of all the data records are then aligned using our partial alignment method. By partial tree alignment, we mean for each pairs of trees (or data records), we only align particular data fields that can be aligned with certainty and ignore those parts that cannot, i.e., making no information on the locations of the unaligned data items. Early uncertain commitments can be result in undesirable effects for later alignment involving other data records. This method turns out to be very effective for multiple tree alignment.

The resulting alignment enables us to extracting data items from all data records in the page. It can be also serves as an extraction pattern to be used to extract data items from other pages with data records generated using the same template.

Our two-step approach called DEPTA (Data Extraction based Partial Tree Alignment), which is more different from all existing methods, does not make those assumptions made by existing methods
As long as a page contains at least two data records, our system will automatically search tags tree. Our experimental results are using a large number of pages show that the proposed technique is highly effective.

## Existing Method

The proposed method is two way processes which extract accurate information from web documents. Given a page, the method first segments the page to identify each data record without extracting its data items. Specifically, the new method also uses visual cues to find data records. Visual information help the system in two ways method:

a)It enables the system to identifying gaps that separate data records, which helps to segment data records correct manner because the gap within a data record (if any) is typically smaller than that in between data records.

b)The proposed system identifies data records by analyzing HTML tag trees or DOM trees. A straightforward way to build a tag tree is to follow the nested tag structure in the HTML code. However, sophisticated analysis has to be incorporated to handle errors in the HTML code (e.g., missing or ill-formatted tags). Whereas the visual or display information can be obtained after the HTML code is rendered by a Web browser, it also contains information about the hierarchical structure of the tags.

In the existing paper, the major web data extraction tools or algorithms are reviewed and compare with each other. This

survey aims at providing a structured and comprehensive overview of the literature in the field of Web Data Extraction. The advantages and disadvantages of the existing methods are also discussed in this survey. The proposed method does not make any assumptions. It only is requiring that the page contains more than one data record, which is almost always true for pages with data records. Our technique contains two steps: (1) identifying data records without extracting each data field in the data records, and (2) aligning corresponding data fields from multiple data records to extract data from them to put in a database table.

## 3. Proposed Method

In this the proposed focuses on the first step: segmenting the Web page to identify individual data records. It does not align or extracting data items in the data records. Since this step is an improvement to our previous technique MDR, algorithm and present the enhancements made to MDR in this work.

### 3.1 MDR Algorithm

The MDR algorithm is based on two observations about data records in a Web page and an edit distance string matching algorithm [2] to finding data records. The two information's are:

1. A group of data records that contains descriptions of a set of similar objects are typically presented in a contiguous region of a page and are formatted using similar HTML tags. Such a region is called a data record region (or data region in short).

The problem with this approach is that the computation is prohibitive because a data record can start from any tag and end at any tags. A set of data records typically does not have same length in terms of its tags strings because it may not contain exactly the same pieces of information. The next observation helps to deal with this problem.

2. The nested structure of HTML tags in a Web page naturally forms a tag tree. In this tree, each data record is wrapped in TR nodes with their sub-trees under the same parent T BODY. The two data records are in the two dash-lined boxes. Our second information is that a set of similar data records are formed by some child sub-trees of the same parent node. It is unlikely that a data record starts in the middle of a child sub-tree and ends in the middle of another child sub-tree. Instead, it starts from the beginning of a child sub-tree and ends at the end of the same or a later child sub-tree.
This information makes it possible to design a very efficient algorithm based on edit distance string comparison to identify data records because it limits the tags from which a data record may start and end in a tag tree.

1. Experimentation show that these information's work very well. We assume that a Web page has only one data region that contains data records. In fact, a Web page may contain a many data regions. Different regions may have different type of data records.
Given a Web page, the algorithm works in three steps

Step 1: Building a HTML tag tree of the page. In the new system, visual (rendering) information is used to build the tag tree.

Step 2: Mining data regions in the page using the tags tree. A data region is an area in the page that contains a list of similar data records. Mining data records directly, which is hard,MDR mines data regions first and then finds data records within seconds. In our new system, again visual information is used in this step to produced better results.

Step 3: Identifying data records from each data region. For example, in Figure 2, this step finds data record 1 and data record 2 in the data region below node TBODY. The main enhancement to the MDR algorithm is the use of visual information to help building more robust trees and also to find more accurate data regions.

### 3.2 Building a HTML Tag Tree

In a Web browser, each HTML element (consisting of a initial tag, opt attributes, opt embedded HTML content, and an finial tag that may be omitted) is denoted in shape of rectangle. A tag tree can be build based on the nested rectangles (resulted from nested tags). The information are as follows:

1. Find the 4 boundaries of the rectangle of each HTML element by calling the embedded parsing and rendering engine of a browser.

2. Detect the containment relationship among the rectangles, A tree can be developed based on the containment check.

This step mines every data region in a page that contains similar data records. Apart of mining data records directly, which is hard, we first mine data regions. By matching tag strings of individual nodes (including their descendents) and combination of multiple adjacent nodes, we can find each data region. We find that nodes 5 and 6 are similar (based on edit distance) and form the data region labeled 1, nodes 8, 9 and 10 are similar and form the data region labeled 2, and the pairs of nodes (14, 15), (16, 17) and(18, 19) are similar and form the data region labeled 3. To avoid using both individual nodes and node combinations, we use the concept of the generalized node to denote each similar individual(tag) node and each (tag) node combination. Thus, a sequence of adjacent generalized nodes forms a data region. Each shaded individual node or node combination in Figure 5 is a generalized node.

### 3.3 Identifying Data Records

After all data regions are found, we find the data records from generalized nodes. We note that each generalized node (a single or a combination of tag nodes in the tag tree) may not represent a single data record. The case can be quite complex. Below, we only highlight two interesting cases in which a data record is not contained in a contiguous segment of the HTML code in order to show some advanced capabilities of our system

Non-contiguous Data Records

In some Web pages, the description of an object (a data record) is not in a contiguous segment of the HTML code. There are two main cases. Figure 6 shows an example of the first case. In this example, the data region contains two generalized nodes, and each generalized node contains two tag nodes (two rows),which indicates that these two tag nodes (rows) are not similar to one and one. But each tag node has the equal number of and the children nodes are similar to one other. One row lists the names of the two objects in two cells, and the next row lists the other pieces of information of the objects also in two cells.

## 4. EXPERIMENTAL RESULTS

In this paper, the traditional Precision, Recall and F-measure will be used to evaluating the experimentation results. The experimentation results for the proposed method for vision-based deep web data extraction for web document clustering are-presented in this section. The proposed approach has been implemented in java (jdk1.6) and the experimentation is performed on a3.0 GHz Pentium PC machine with 2 GB main memory. For experimentation, we have taken more deep web pages which contained all the noises such as Navigation bars, Panels and Frames, Other Uninteresting Data. These pages are then applied to the proposed method for removing the different noises. The removal of noise blocks and extracting of useful content chunks are explained in this sub-section.

GDS: Our data set is collected from the complete planet web site. Complete-planet is currently biggest depository for search entries of more than 70,000 web databases and search engines. This Web database are classifying into 42 categories covering most domains in the real world. GDS consists of 1,000 available Web databases. For each Web database, we substituting many queries and gather five deep Web pages with each containing at least three data records.

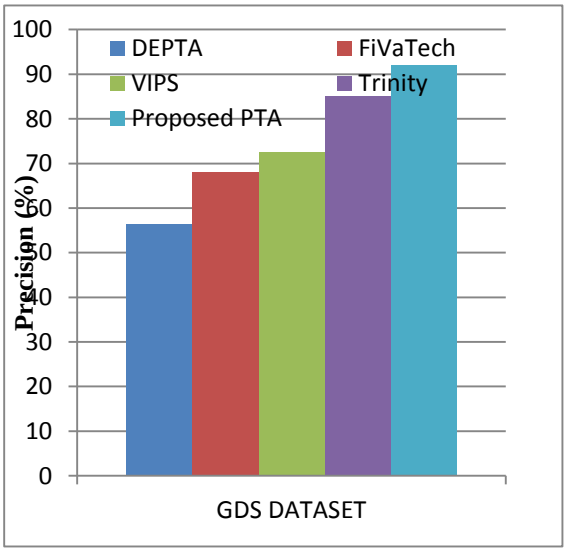


Figure 4.1

Evaluation result of precision recall and F-measure

## 5. CONCLUSION

In this paper, the major web data extraction tools or algorithms are reviewed and compared with each other. This survey providing a structured and comprehensive overview of the literature in the field of Web Data Extraction. The advantages and disadvantages of the existing methods are also discussed in this survey. A new approach is also proposed in this survey to extract structured data from Web pages. It only requiring that the page contains more than one data record, which is almost always true for pages with data records. Our technique contains of two steps: (1) identifying data records without extracting each data field in the data records, and (2) aligning particular data fields from multiple data records to extract data from them to put in a database table.

## REFERENCES


1. Emilio Ferraraa, Pasquale De Meob, GiacomoFiumarac , Robert Baumgartnerd, "Web data extraction, applications and techniques: A survey",Knowledge-Based Systems Volume 70, November 2014, Pages 301–323.
2. Chia-Hui Chang, Mohammed Kayed, MohebRamzyGirgis, KhaledShaalan"A Survey of Web Information Extraction Systems",IEEE Transactions On Knowledge And Data Engineering, vol.18, no.10, pp.1-18.
3. M. Newman. The structure and function of complex networks. SIAM review, pages 167–256, 2003
4. L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. Arxiv preprint arXiv:1111.4570, 2011.
5. Fang Dong, Mengchi Liu and Yifeng Li, "Automatic Extraction of Semi-structured Web Data",International Journal of Database Theory and Application,Vol. 6, No. 4, August, 2013.
6. Tim Furche, Georg Gottlob, Giovanni Grasso, Christian Schallhart, Andrew Sellers, "OXPath: A language for scalable data extraction, automation, and crawling on the deep web" The VLDB Journal February 2013, Volume 22, Issue 1, pp 47-72
7. Yong Feng,DongfengJia ,Huijuanwang, "PFIME: Parallel Automatic Deep Web Data Extraction Based on Hadoop", Journal of Computational Information Systems vol.10, no.9 pp.3863–3870, 2014.
8. Srikantaiah K.C., Suraj M., Venugopal K.R., Iyengar S.S., and L.M. Patnaik, "Similarity Based Web Data Extraction and Integration System for Web Content Mining"Advances in Communication, Network, and Computing Lecture Notes of the Institute for Computer

Sciences, Social Informatics and Telecommunications Engineering Volume 108, 2012, pp 269-274.

9. Kayed, Mohammed; Chia HuiChang ;Shaalan, K. ; Girgis, M.R. "FiVaTech: Page-Level Web Data Extraction from Template Pages",Seventh IEEE International Conference on Data Mining Workshops, pp.15 - 20 ,2007.
10. D. Raghu ,V. Sridhar Reddy ,Ch. Raja Jacob ,"Dynamic Vision-Based Approach in Web Data Extraction",International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, 2734-2736.
11. Wei Liu, XiaofengMeng, and WeiyiMeng "ViDE: A Vision-based Approach for Deep Web Data Extraction",IEEE Transactions on Knowledge and Data Engineering, Volume:22 , Issue: 3,pp.447 – 460,2009.
12. M. A. Kaufmann,E. Portmann, M. Fathi , "A Concept of Semantics Extraction from Web Data by Induction of Fuzzy Ontologies" , IEEE International Conference on Electro/Information Technology (EIT), pp.1-6, 2013.
13. Cheng Cui "Heterogeneous Web Data Extraction Algorithm Based On Modified Hidden Conditional Random Fields"Journal Of Networks, Vol. 9, No. 4, pp. 993-999, April 2014.
14. Shengsheng Shi, Wu Wei, Yulong Liu, Haitao Wang, Lei Luo, Chunfeng Yuan, Yihua Huang "NEXIR: A Novel Web Extraction Rule Language toward a Three-Stage Web Data Extraction Model",Web Information Systems Engineering Lecture Notes in Computer Science Volume 8180, 2013, pp 29-42.
15. Jer Lang Hong ; "Deep web data extraction"IEEE International Conference on Systems Man and Cybernetics (SMC), pp.3420 - 3427 ,2010.
16. HaikunHong ;Xiaoxin Chen ; Guoshi Wu ; Jing Li "Web Data Extraction Based on Tree Structure Analysis and Template Generation", International Conference on, E-Product E-Service and E-Entertainment ICEEE pp.1 – 5, 2010.
17. Nitin Jindal and Bing Liu "A Generalized Tree Matching Algorithm Considering Nested Lists for Web Data Extraction"Proceedings of the 2010 SIAM International Conference on Data Mining, pp.930-941.
18. Tak-Lam Wonga, Wai Lamb, "An unsupervised method for joint information extraction and feature mining across different Web sites",Data & Knowledge Engineering Volume 68, Issue 1, January 2009, Pages 107–125.