Web Data Extraction and Alignment Tools: A Survey

Pranali Nikam[1] Yogita Gote[2] Vidhya Ghogare[3] Jyothi Rapalli[4]
[1,2,3,4]Department of Information Technology
[1,2,3,4]Pune University

Abstract— Data extraction from the web pages is the process of analyzing and retrieving relevant data out of the data sources (usually unstructured or poorly structure) in a specific pattern for further processing, involves addition of metadata and data integration details for further process in the data workflow. This survey describes overview of the different web data extraction and data alignment techniques. Extraction techniques are DeLa, DEPTA, ViPER, and ViNT. Data alignment techniques are Pairwise QRR alignment, Holistic alignment, Nested structure processing. Query Result pages are generated by using Web database based on Users Query. The data from these query result pages should be automatically extracted which is very important for many applications, such as data integration, which are needed to cooperate with multiple web databases. New method is proposed for data extraction t that combines both tag and value similarity. It automatically extracts data from query result pages by first identifying and segmenting the query result records (QRRs) in the query result pages and then aligning the segmented QRRs into a table. In which the data values from the same attribute are put into the same column. Data region identification method identify the noncontiguous QRRs that have the same parents according to their tag similarities. Specifically, we propose new techniques to handle the case when the QRRs are not contiguous, which may be due to presence of auxiliary information, such as a comment, recommendation or advertisement, and for handling any nested structure that may exist in the QRRs.

Key words: Combining Tag And Value Similarity (CTVS), Query Result Record (QRR), Data extraction and label assignment for web database (DeLa), Data Extraction Based on Partial Tree Alignment (DEPTA), Visual Perception based Extraction of Records (ViPER), Visual information and Tag Structure based wrapper generator (ViNTS)

I. INTRODUCTION

Online database comprises the deep web, also known as web databases. Data are usually embedded in html pages. Automatic extractions of data after a query input from web are carried out with the help of techniques such as Wrapper system and automatic extraction. Extracted data if organized in a structured manner, such as tables, so that they can be compared and aggregated. Accurate data extraction is essential for any application to execute correctly. The main objective of this paper is to extract data automatically from multiple web databases and align them in one format.

Many web applications such as data integration, shopping sites comparison and meta-querying, need the data from multiple databases. In general, when a query is placed, result pages not only contain the actual data instead it involves other information too, such as navigational panels, comments, various advertisements and links of various sites. Extracting only the required and efficient data is essential for any applications growth. Combining Tag and value Similarity (CTVS) is a two-step method, to extract Query Result Record (QRR) from any query result page which is called as p.

1) Extraction of Records: Identifies the QRRs in the page p and performs two steps which is data region identification and actual segmentation step.
2) Alignment of record: Aligning the value to the data of the QRRs in the page p into a structured format such as tables which can be further used for aligning the same attributes into the same table column.

This paper focuses on aligning the QRR which can be performed in three steps which are Pair-wise QRR alignment, Holistic alignment and Nested structure processing. These develop a technique to allow QRRs to be contiguous in data region.

Paper deals with the study of automatic web data extraction and data alignment technique. Wed data extraction technique are mainly classified as

1) Wrapper Programming Language: This uses special pattern with specification language that help user for development extraction program
2) Wrapper Induction Method: A technique for construction of wrapper automatically labeled of the resource content. Approach uses mainly extraction a rule which is derived from inductive learning. A user label the present item form set of present training pages or from records on the page, which we want to be extracted as tangent item form the page. The system which learns wrapper rule form the label and then uses for extraction record form new pages

A. Rules:

1) Prefix pattern: Denote beginning of a target item.
2) Suffix pattern: Denotes end of a target item.
3) Automatic Extraction Methods: Tag structure is used on query result pages. Used form the wrapper induction problems the unsupervised learning methods are comes to know which deals with automatic extraction method.

1) Advantages:

User itself labels the items in which he was interested so that there is no available extra data is given to him or extracted from the database.

2) Disadvantages:

1) Time consuming as we are using the manual labeling and further used for large database as system is not scalable.
2) Need to keep track of the page format, continuous monitoring is needed hence existing wrapper gives poor performance

II. DELA (DATA EXTRACTION AND LABEL ASSIGNMENT FOR WEB DATABASE)

DeLa system automatically extract data from website and associate meaningful labeling to this data .It mainly uses

keyword to keep track on page DeLa uses to fetch the text from the web page into the table and assign label with the help of four component (fig1) form crawler ,wrapper generator, data aligner ,label assigner. DeLa system that sends queries through HTML forms, automatically extracts data objects from the retrieved web pages, fits the extracted data into a table and then assigns labels to the attributes of the data objects, which the columns of the table. Based on two main observations. First, data objects contained in dynamically generated web pages share a common HTML tag structure and they are listed continuously in the web pages. Based on this, we automatically generate regular expression wrappers to extract such data and fit them into a table. Second, the form contained in a web page, through which users submit their queries then we extract the labels of the HTML form elements and match them to the columns of the data table, thereby annotating the attributes of the extracted data. The DeLa system consists of four components: a form crawler, a wrapper generator, a data aligner and a label assigner. First component, we adopt the existing hidden web crawler and send queries to the web site. In the remaining three components, we make the following contributions. First, we develop a method to automatically generate regular expression wrappers from data contained in web pages. Second, we employ a new data structure to record the data extracted by the wrappers and develop an algorithm to fill a data table using this structure. Third, we explore the feasibility of heuristic-based automatic data annotation for web databases and evaluate the effectiveness of the proposed heuristics.

A. Form Crawler:

Collect label from the web site hidden web crawler use in DELA and make regular expression wrapper then sent it to wrapper generator.

B. Wrapper Generation:

The input comes from form generator make final regular expression base on HTML tag structure of page if page contains more than one instance of the same data object that tag enclosing object may appear repeatedly. It considers each page as sequence of taken made up of HTML tag.
Example: Special token "text" used to represent the text string

C. Data Alignment:

1) Data Extraction

This step does the data extraction from web pages by using the wrapper produced by wrapper generator. Then loaded into table to represent web page mainly use regular expression pattern and token sequence data is analyzed and crawled through to retrieve relevant information from data sources like a database in a specific pattern. Further processing is done, it includes adding metadata and other data integration; another process in the data workflow is called data extraction.
The data extraction mainly comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.

2) Attribute Separation:

Its prerequisite is removal of tags. If several attribute are appearing in one text string then they must be separated by symbol.

D. Label Assigner:

Its heuristic match, search, encoded search and data attribute conventional formats.

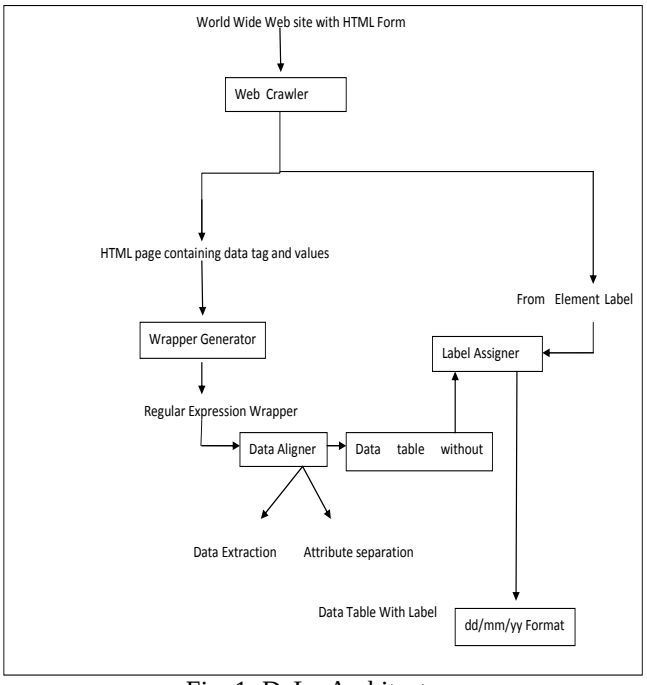

Fig. 1: DeLa Architecture

1) Advantages:

1) It requires the edge between one internal node with one leaf node which has a label and that will be is the token at the position of the internal node in the suffix starting from the leaf node.
2) Alphabetical order such as sibling nodes sharing the same parent are put in same order.

2) Disadvantages:

1) It requires the two sub-steps are Production of one rooted tag tree for each data record with sub-tree all data arranged in single tree.
2) No data item is involved in the matching process it requires only tag nodes for matching.
3) Sequence of token and produces a rules or regular expression for each page hence sequencing is needed for each page.

III. DEPTA (DATA EXTRACTION BASED ON PARTIAL TREE ALIGNMENT)

The first step of DEPTA also called MDR-2, with existing system MDR for identifying data records.MDR is already more effective than them. Data items/fields alignment and extraction It is the second step of DEPTA which is able to perform the same task. It performs poorly in finding right data records, and thus could not extract data items well. Contain similar data records to find patterns from the pages to extract data items. The technique in the experiments, such detail pages are manually identified and downloaded, which is unrealistic in practice. DEPTA is more general. Given a single page, it is able to extract data records and a data item from it.DEPTA system performs automatic data extraction given a single page with lists of records of data (fig 2).

A. Component:

Building HTML tag tree (DOM tree).it detect containment relationship between rectangle of each HTML tag

B. Mining Data Region:

Find the data region by comparing tag string associated with individual node with combination of multiple adjacent and descendants gaps present between data record are used eliminate false node combination.

C. Identifying Data Records:

The data records are not contiguous in two cases. The node has same number of children and two or more data regions from multiple data records.

D. Data Item Extractor:

Sub tree of all data arranged into single tree and uses partial alignment which was based on the tree matching. For matching only tag nodes are used. Seed tree is a tree with a minimum number of data fields which is at its initial. The selection of a seed tree is that it has good alignment with data fields in other data records for each treeTi [i! =s] the algorithm tries to find out the match node Ts. When match found for node ni link between ni to ns. If no match found then algorithm attempts to expand seed tree by insert ni tons.

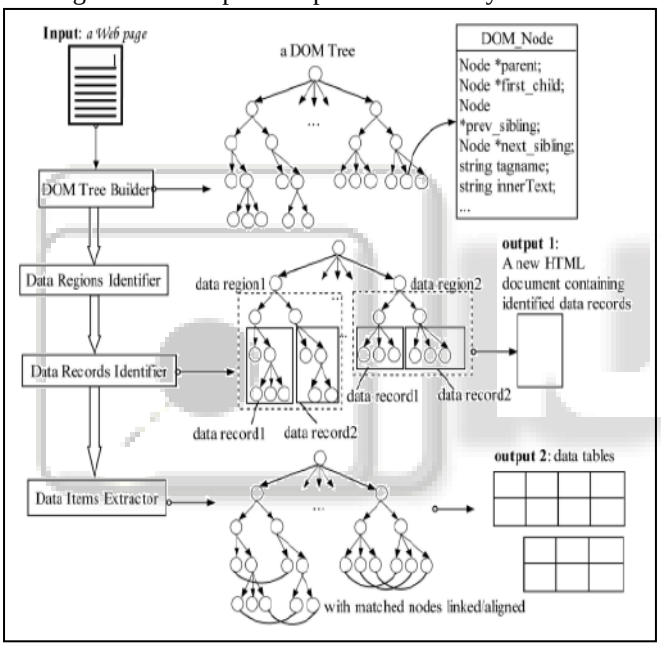

Fig. 2: DEPTA Architecture

1) Advantages:

1) Extraction is performed mainly by partial tree alignment structure.
2) DEPTA will find out nested records in addition to flat records. All remaining techniques produce only flat records.

2) Disadvantages:

1) Manual labeling is time-consuming and it is hard to scale to a large number of sites on the one web page.
2) recall and precision are computed based on the basis of total number of correct data records found in all pages and the actual number of data records in pages
3) Data item extraction of DEPTA, the precision and recall computation has considered the 20 lost data records because of step 1 of DEPTA which is not good for extraction process.

IV. VIPER (VISUAL PERCEPTION BASED EXTRACTION OF RECORDS)

A new fully-automatic information extraction tool named ViPER The principle is that a Web page contains at least two multiple consecutive data records building a data region which have exhibits some kind of structural and visible similarity. ViPER is able to extract and discriminate repetitive information contents with respect to the user's visual perception of the Web page. Having identified the most relevant data region the which have not yet been applied, which show the efficiency and accuracy of the extraction and alignment process we used an available third-party test with manually extracted and labeled data. We compared ViPER with existing state-of-the-art wrapping tools. Viper ifs fully automated information extraction tool which works on webpage which contains at least two consecutive records. ViPER extract relevant data with respect to visual perception of web page ViPER uses both visual data value similarity features and HTML tag structure identify repetitive pattern.
Two step process:

A. Data Extraction:

HTML document is labeled unordered tree which (V,E,r,n)

- V=Vertices
- E=Set of edges
- R=Root
- N=Label function n=V*L
- L=String

B. Data Alignment:

Uses general suffix tree5 for every sequence we cannot extends a match to left or right and unique means match occur only in each of n sequences. If we try to align data record using MUM(Maximal Unique Match) sequence of length the 1 then this problem decomposes into 1+1 smaller unaligned sub region.

1) Advantages:

ViPER uses several techniques which to extract relevant information from dynamic Web Sites. The algorithm is able to extract repetitive data records with high precision and our system with current state-of-the art systems on several third party data sets with very good results.

2) Disadvantages:

1) ViPER, used dataset 2 from the ViNTs Web page featuring sample pages from 100 different search engines which is with the huge amount of data.
2) Dataset to compare with ViNTs system with the available MDR tool, calibrating the similarity threshold value at only 60%.

V. VINTS (VISUAL INFORMATION AND TAG STRUCTURE BASED WRAPPER GENERATOR):

It is tool for automatically producing wrapper which extracts search result records (SRR) from dynamically generated HTML result pages returned by any search engine. It first utilizes the visual data value similarity without considering the tag structure to identify data ViNTs learns a wrapper from a set of training pages from a website. It first utilizes the visual data value similarity without considering the tag structure value similarity regularities denoted as data value similarity lines and then combines them with HTML tag.

The input to system is URL of search engines interface page which contains an HTML form used to accept user queries.

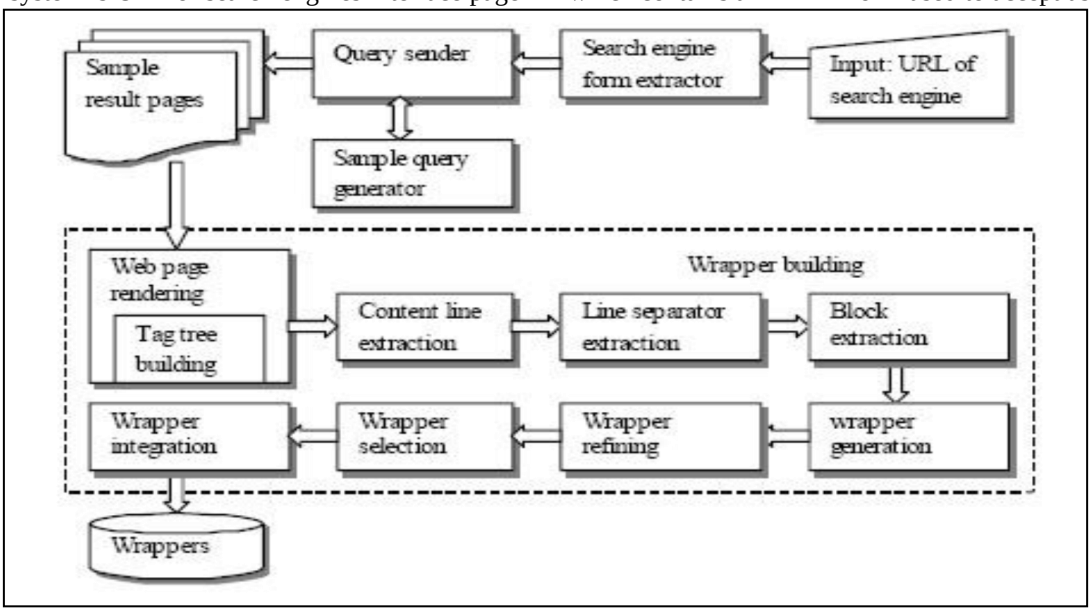

Fig. 3: ViNT Architecture

A. Advantages:

1) Dynamically, have structure data records template generated HTML pages publishing data objects that are stored in back-end databases with them.
2) ViNTs technique used for automatically producing the wrappers for any given search engines.

B. Disadvantages:

1) ViNTs get the wrapper from a set of training pages from a website. It mainly utilizes the visual data value similarity without considering the tag structure only.
2) If the data records are distributed over multiple data regions then only major data region are reported.
3) It requires users to collect the training pages from the website including the no result page which may not exist for many present web databases because they respond with records that are close to the query if no record matches the query exactly.
4) The pre-learned wrapper usually fails when the format of the query result page changes. Hence it is necessary for ViNTs to monitor that is continuous monitoring of format changes to the query result pages which are most difficult problem.

VI. RELATED WORK

DeLa models the structured data contained in template-generated webpage as string instances encoded in HTML tags, of the implied nested type of their web database. A regular expression is employed to model the HTML-encoded version of the nested type. Since the HTML tag-structure enclosing the data may appear repeatedly if the page contains more than one instance of the data, the page is first transformed into a token sequence composed of HTML tags and a special token "text" representing any text string enclosed by pairs of HTML tags. Then, continuous repeated substrings are extracted from the token sequence and a regular expression wrapper is induced from the repeated substrings according to some hierarchical relation-ships among them. The main problem with this method is that it often produces multiple patterns (rules) and it is hard to decide which is correct.

Deriving accurate wrappers based solely on HTML tags is very difficult for the following reasons.

- One cannot rely on "proper" HTML tag usage since HTML tags are often used in unexpected and unconventional ways.
- HTML tags convey little semantic information since their main purpose is to facilitate the rendering of data.
- Some data may contain embedded tags, which may confuse the wrapper generators making them even less reliable.

To overcome these shortcomings, methods such as ViPER and ViNTs make use of additional information in the query result pages.

ViPER uses both visual data value similarity features and the HTML tag structure to first identify and rank potential repetitive patterns. Then, matching subsequences are aligned with global matching information. While ViPER suffers from poor results for nested structured data, CTVS handles nested structured data efficiently and precisely.

Using both visual and tag features ViNTs learns a wrapper from a set of training pages from a website. It first utilizes the visual data value similarity without considering the tag structure to identify data value similarity regularities, denoted as data value similarity lines, and then combines them with the HTML tag structure regularities to generate wrappers. Both visual and nonvisual features are used to weight the relevance of different extraction rules. Several result pages, each of which must contain at least four QRRs, and one no-result page are required to build a wrapper. The resulting wrapper is represented by a regular expression of alternative horizontal separator tags (i.e., <HR> or

), which segment descendants into QRRs.

ViNTs has several drawbacks. First, if the data records are distributed over multiple data regions only the major data region is reported. Second, it requires users to

collect the training pages from the website including the no-result page, which may not exist for many web databases because they respond with records that are close to the query if no record matches the query exactly. Third, the prelearned wrapper usually fails when the format of the query result page changes. Hence, it is necessary for ViNTs to monitor format changes to the query result pages, which is a difficult problem. In contrast, CTVS requires neither training pages nor a prelearned wrapper for a website. However, unlike ViNTs, CTVS cannot handle no-result pages, since CTVS assumes there are at least two QRRs in the page to be extracted.

All the preceding works make use of only the information in the query result pages to perform the data extraction. There are works that make use of additional information, specifically ontologies, to assist in the data extraction. While these approaches can overcome some of the limitations of CTVS (e.g., that a query result page must contain at least two QRRs) and can achieve high accuracy, they require the availability of additional re-sources to construct an ontology as well as the additional step of actually constructing the ontology

	Nested Structure Processing	Single Result Page	Non-contiguous Data Regions
CTVS	√	√	√
DeLa	√	√	×
ViPER	×	√	×
ViNTs	×	×	×

Table 1: Data Extraction Method Summarization

Table summarizes some characteristics of the data extraction methods compared in this paper. The single page result column indicates whether a single query result page from a data source is sufficient to extract data.

VII. Conclusion

The Existing Data Extraction Method allows the Query Result Records in a data region to be non-contiguous as well as aligns the data values among the QRRs. Although it has been shown to be an accurate data extraction method it does not figure out the case where multiple data values from more than one attribute are clustered inside one leaf node of the tag tree and data value of a single attribute spans multiple leaf nodes. Here the proposed structural-semantic entropy is calculated for each node in a DOM tree. It focus on recognizing data-rich regions and find the lowest common parent nodes of the sibling sub trees forming the records in the DOM tree representation of a web page with the help of a set of domain keywords. The future work may be extended to extract the data from webpage based on the design issues such as memory consumption, computational overhead, storage, fast processing etc. The algorithm used requires that the entropy should be calculated for every non-leaf node of a DOM tree. One of the possible approach is to find rules to terminate the calculation before the entropies of all nodes are calculated in a bottom-up way. On the other hand accelerate data extraction for the pages during the process of crawling a web site.

VIII. Future Enhancements

Page ranking method has been introduced based on giving the weight to the particular web page for speeding up the extraction of web page process. Clustering algorithm is to stipulate the clustering of related web pages so as to reduce the collision that has been taken place in the previous CTVS method. We improved our algorithm with these existing techniques by allowing the QRRs in a data region to be non-contiguous. A novel alignment method is proposed in which the alignment is performed in three consecutive steps: pair wise alignment, holistic alignment, and nested structure processing. Experiments on five sets show that CTVS is generally more accurate than current state-of-the-art methods. Although CTVS has been shown to be an accurate data extraction method, it still suffers from some limitations. First, it requires at least two QRRs in the query result page. Future enhancements of page ranking method and clustering methods and the time consumption in retrieving the data has been reduced accordingly Website linking structure has been identified and implemented in order to find the linkage between the web pages. Future enhancement as an application, the page ranking method or algorithm has been implemented also if a campaign of exchanging links to increase Page Rank is to be implemented, it is vital that the importance of factors such as link text is understood. Page Rank declares that a link from a rarely visited and rarely updated web page should not have equal weighting to a link from a popular web page. The purpose of the Page Rank algorithm is to attach a score, ranging from zero to ten, to every web page.

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, " VOL 24,NO7,JULY 2012

[2] Shridevi A. Swami,"Web Data Extraction and Alignment Tools: A Survey", PujashreeVidap Department of Computer Engineering, Pune Institute of Computer Engineering Pune, India. 2003

[3] Anuradha R. Kale, Prof V.T.Gaikwaid, Prof H.N.Datir,"Data Extraction and alignment for multiple web Databases",International Journal of Scientific & Engineering Research, Volume 4, Issue 7, July-2013 ISSN 2229-5518

[4] A. Arasu and H. Garcia-Molina,"Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Intl Conf. Management of Data, pp. 337-348, 2003.

[5] R. Baeza-Yates, "Algorithms for String Matching: A Survey," ACM SIGIR Forum, vol. 23, nos. 3/4, pp. 34-58, 1989.

[6] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual Web Information Extraction with Lixto," Proc. 27th Intl Conf. Very Large Data Bases, pp. 119-128, 2001.

[7] M.K. Bergman,"The Deep Web: Surfacing Hidden Value," White Paper, BrightPlanet Corporation,

http://www.brightplanet. com/ re-sources/details/deepweb.html, 2001.

[8] P. Bonizzoni and G.D. Vedova,"The Complexity of Multiple Sequence Alignment with SP-Score that Is a Metric," Theoretical Computer Sci-ence, vol. 259, nos. 1/2, pp. 63-79, 2001.

