Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Data Extraction of a Novel Method for Clustering Alignment based on Combining Tag and Value Similarity

S.Suganya M.E, R.Devi M.E
[12] ***Assistant*** *Professor, Department of CSE,*
R.V.S Educational Trust's Group of Institutions,
Dindigul, Tamilnadu, India

T.Thangam M.E
Associate Professor, Department of ECE,
P.S.N.A College of Engineering and Technology,
Dindigul, Tamilnadu, India

Abstract: Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Data Extraction is the process of retrieving data out of data sources. Establishment of an updated method called a novel data extraction and alignment method called CTVS that combines both tag and value similarity enhances the efficiency of the data extraction and alignment. Record alignment algorithm has been introduced in order to perform efficient contemporary alignment method first pair wise and then holistically. Threshold index formula has also been introduced to find the data regions and to perform clustering methods. The applications of this extraction of data records also includes the concept of page ranking in order to speed up the search engine, clustering the similar data regions in order to perform efficient identification of web pages and similar Query result pages and applicable also in data integration and comparison shopping.

Keywords: Data Extraction, automatic wrapper generation, data record alignment, information integration, page ranking, clustering method.

I. INTRODUCTION

Web databases comprises deep web from these web sites and web pages the extraction of the needed data from a web page is merely a complicated process. If a use user gives a query it automatically extracts the query result pages.

Many web applications, such as Meta querying, data integration and comparison shopping, need the data from multiple web databases. For these applications to further utilize the data embedded in HTML pages, automatic data extraction is necessary. Only when the data are extracted and organized in a structured manner, such as tables, can they be compared and aggregated. Hence, accurate data extraction is vital for these applications to perform correctly. This paper focuses on the problem of automatically extracting data records that are encoded in the query result pages generated by web databases. In general, a query result page contains not only the actual data, but also other information, such as navigational panels, advertisements, comments, information about hosting sites, and so on.

We employ the following two-step method, called Combining Tag and Value Similarity (CTVS), to extract the QRRs from a query result page p.

1. **Record extraction process** identifies the QRRs in p and involves two sub steps: Data region Identification and the actual segmentation step.
2. **Record alignment process** aligns the data values of the QRRs in pinto a table so that data values for the same attribute are aligned into the same table column. Compared with existing data extraction methods, CTVS improves data extraction accuracy in three ways.

1. Efficiency of the new techniques are proposed to handle the case when the QRRs are not contiguous in p, which may be due to the presence of auxiliary information. Assume that the QRRs are presented contiguously in only one data region in a page.

a. An adapted data region identification method is proposed to identify the non-contiguous QRRs that have the same parents according to their tag similarities.

b. A merge method is proposed to combine different data regions that contain the QRRs (with or without the same parent) into a single data region. Our experimental results show that the two techniques are effective for addressing the non-contiguous data region problem.

2. A contemporary alignment algorithm has been introduces in order that it provides an efficient aligning process, first pair wise then holistically, so that they can be put into a table with the data values belonging to the same attribute arranged into the same table column.

3. A new nested-structure processing algorithm is proposed to handle any nested structure in the QRRs after the holistic alignment. Unlike existing nested-structure processing algorithms that rely on only tag information, the updated CTVS method uses both tag and data value similarity information to improve nested-structure processing accuracy.

II. OVERVIEW

In this paper the problem is focused on extracting the data from web pages and alignment of data. The goal of web database data extraction is to remove any irrelevant information from the query result page, extract the query result records from the page, and align the extracted QRRs into a table such that the data values belonging to the same attribute are placed into the same table column. In order to enhance an efficient data extraction method in order to reduce the time consumption of extraction the data two methods have been introduced.

2.1 PAGE RANKING

Page Rank is an important part of search engine optimization (SEO) past and present history. Page Rank is a link analysis algorithm that assigns a number or rank to each hyperlinked web page within the World Wide Web. The basic purpose of Page Rank is to list web pages from the most important to the least important, reflecting on a search engine results page when a keyword search occurs. The basic process involves Page Rank evaluating all of the links to a particular web page. If a web page has a lot of links from large websites that also rank well, then the original web page is given a high ranking. Higher ranking in Page Rank equates to a greater probability of the site being reached because they are not only quantitatively linked to a number of times, but are also linked to other popular web pages. We have also used page ranking method for speeding up the search engine in order to extract the data records rapidly in order to reduce the extraction time of the data records.

2.2 CLUSTERING OF DATA REGIONS

Finding the group of objects such that the objects in a group will be similar or related to one another and different groups from the objects will be put into another group. In an existing CTVS method Page ranking and clustering algorithms have not been introduced. Enhancement of these algorithms will give an extra value to the web pages as per the page ranking algorithm and similar pages will be clustered in order to extract the needed information immediately which reduces the
time consumption.

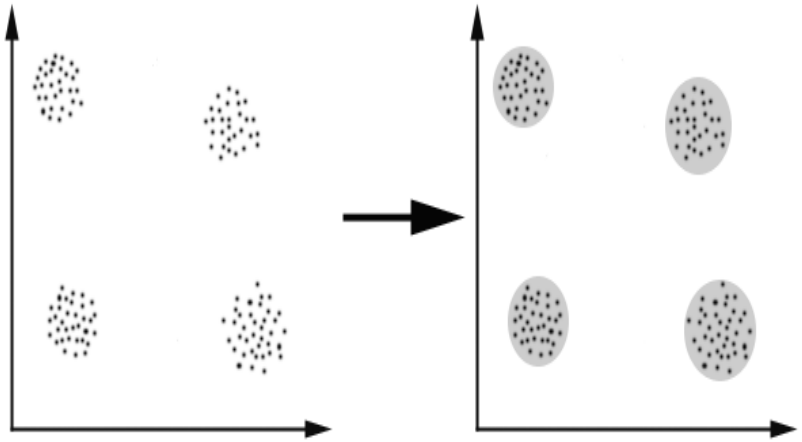

Fig 1. Clustering Methods in Data Mining

III. DESIGN CONSIDERATION

Design consideration of this project includes a novel data extraction method, CTVS, to automatically extract QRRs from a query result page.

Tag tree construction module has been implemented in order to check from the parent root tag and traversing along the child node path and reaches the bottom most tag or node. By having similar tags and also the similarity between the values of two data regions have been obtained and arrangements of those values have implemented to perform effective alignment or aggregation method.

3.1 QUERY PAGES AND QRR EXTRACTION

Each node represents a tag in the HTML page and its children are tags enclosed inside it. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions.

Overcoming the problems of existing system this method involves in combining the tags and it form a data region of similar tag and in parallel it will also compare the value similarity of the data regions to find out the similar values in order to align it in to the table.

3.1.1 Data mining and Region Identification

In this data region method we first assume that some child sub trees of the same parent node form similar data records, which assemble a data region. The similar data records are typically clustered using several clustering.

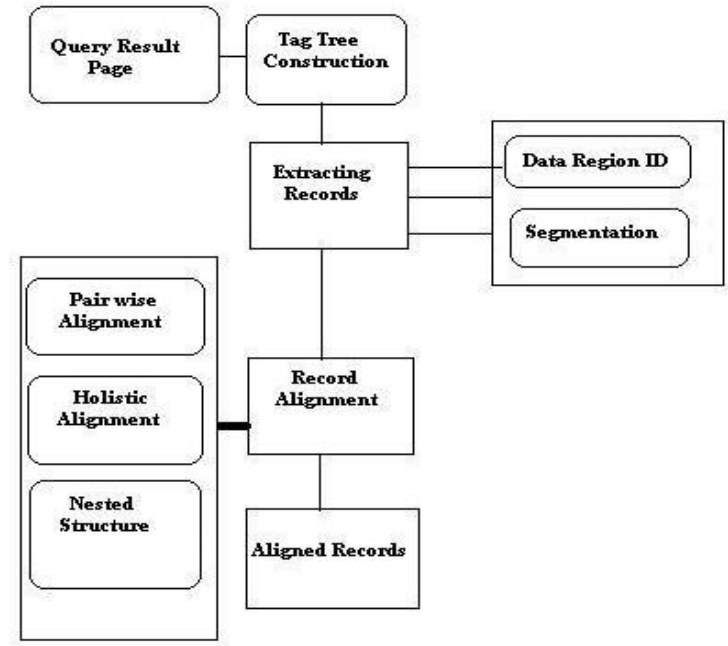

Fig. 2 QRR Extraction Frame work

3.1.2 Web page record Segmentation

To illustrate the record segmentation algorithm, assume that we have Region 1 and Region 2 we have to find out the tandem repeats in order to know there is a record for example ABABABA if we use characters A to represent an element of the similar node set and B to represent and element of the similar another node set In this case, there are two tandem repeats, AB and BA.

3.1.3 Identification of Data Region Merge

After the clustering methods have been introduced we need to determine whether any of the data regions should be merged.

Given any two data regions, we treat them as similar if the segmented records they contain are similar. The similarity between any two records from two data regions is measured by the similarity of their tag strings. The similarity between two data regions is calculated as the average record similarity. Two data regions can be merged into a merged data region if the records in the two data regions have an average similarity greater or equal to 0.6, which is a threshold used to judge whether two records are similar in regions.

TABLE 1: Final Aligned Result

QRR 1	Image 1	DNK Y	Siz e:	Quee n	Color :	Blue	$55
QRR 2	Image 2	DNK Y	Siz e:	Quee n	Color :	Red	$56
QRR 2	Image 2	DNK Y	Siz e:	King	Color :	Gree n	$57

3.1.4 Query Result Section Identification

Even after performing the data region merge step, there may still be multiple data regions in a query result page.

1. The query result section usually occupies a large space in the query result page.
2. The query result section is usually located at the center of the query result page.
3. Each QRR usually contains more raw data strings than the raw data strings in other sections. The above three weights are summed and the data region that has the largest summed weight is selected as the query result section. Records in this data region are assumed to be QRRs.

3.2 ALINMENT OF QUERY RESULT RECORD

QRR alignment is performed by a novel three-step data alignment method that combines tag and value similarity. Based on the clustering method we have introduced efficient pair wise, holistic and nested structure methods.

3.2.1 Pair wise QRR Alignment

Same record path constraint. The record path of a data value f comprises the tag from the root of the record to the node that contains f in the tag tree of the query result page. Each pair of matched values should have the same tag path then unique constraint. Each data value can be aligned to at most one data value from the other QRR after the no cross alignment constraint will be checked.

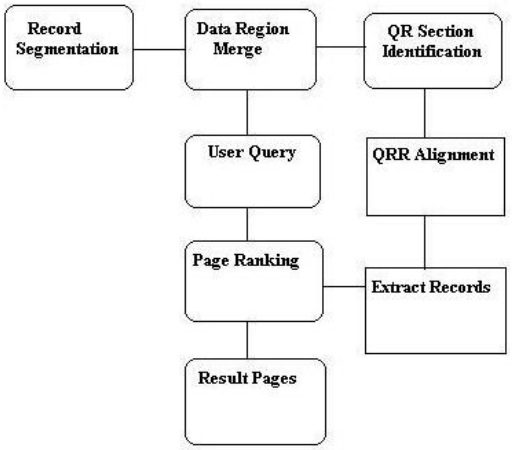

Fig 3. Frame work of QRR Extraction and Page

TABLE 3: DATA EXTRACTION METHOD

Method	**Nested Structure Processing**	**Single Result Page**	**Non Contiguous Data Regions**
CTVS	Yes	Yes	Yes
DeLa	Yes	Yes	No
ViNTs	No	No	No

According to (1), the process of finding the alignment with the largest sum of similarity is shown in Table 3. Starting from the alignment betweenf11andf21, the cells in bold italic represent the identified alignment between the data values.

3.2.1.1 Data Value Similarity Calculation

Given two data valuesf1 andf2 from different QRRs, we require their similarity, s12, to be a real value in [0, 1]. The data value similarity is calculated according to the data type tree is shown in Table 5.Each child node is a subset of its parent node.

TABLE 2
Example: Comparison and Data Alignment

Scientific Company	Name of the instrument	Manufacturer	Price
United Scientific Company	Travelling Microscope	Besto	7,200
Modern Scientific	Travelling Microscope	Besto	6,000
SLS Scientific	Travelling Microscope	Besto	6,500

Fig 4. Tag Tree Construction

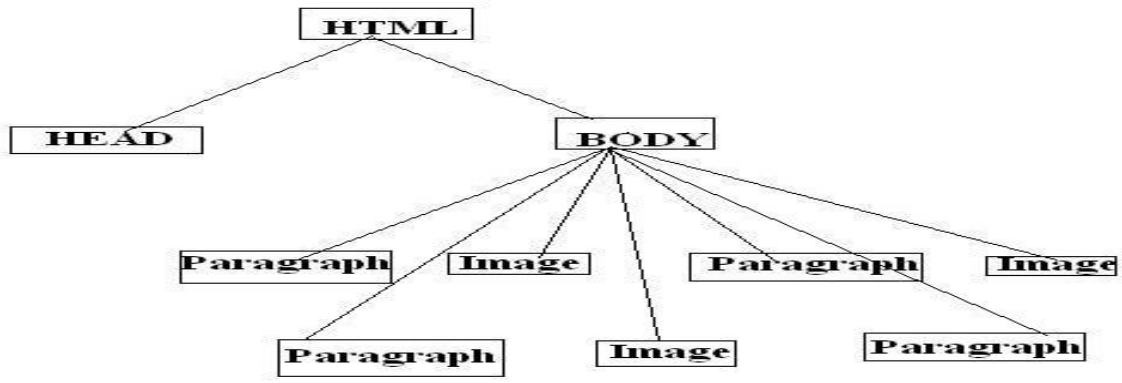

Fig 3. Frame work of QRR Extraction and Page

Given two data valuesf1andf2, we first judge their data types and then fit them as deeply as possible into the nodesn1 andn2 of the data type tree. For example, given a string"784," we will put it in node "integer."

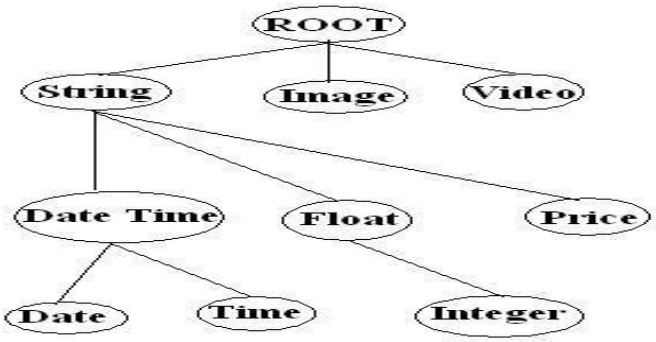

Fig 5. Data Type Tree

3.2.2 Break Path and Holistic Alignment

Given the pair wise data value alignments between every pair of QRRs, the step of holistic alignment performs the alignment globally among all QRRs to construct a table in which all data values of the same attribute are aligned in the same table column. To consider two application constraints that is specific to our holistic alignment problem.

3.2.3 Nested Structure Processing

Holistic data value alignment constrains a data value in a QRR to be aligned to at most one data value from another QRR. If a QRR contains a nested structure such that an attribute has multiple values, then some of the values may not be aligned to any other values. Therefore, nested structure processing identifies the data values of a QRR that are generated by nested structures (i.e., the repetitive parts of a generating template).

1. In an existing CTVS processes the nested structures after the data records are aligned rather than before as is the case in DeLa and NET. But due to the implement of new page ranking and clustering methods the above problems will be solved accordingly.
2. In an Updated CTVS the data value similarity information effectively prevents a flat structure from being identified as a nested structure. Because it shares similar tag structures, a flat structure with several columns having the same tag structure, might be mistakenly identified as a nested structure in DeLa and NET. Incorrectly identifying a flat structure as a nested one can have serious consequences.

IV. EVALUATION SETUP

We have performed different tests to assess the performance of updated CTVS algorithm. We now present the experimental results for updated CTVS over five data sets and compare Updated CTVS with ViNTs, and DeLa. We have chosen ViNTs and DeLa to compare with CTVS because both have been shown to perform very accurate data extraction and implementations of both are available to us. An Updated CTVS is implemented in JAVA and C++. When running on a Pentium 4 2.8 GHz CPU with 1 GB memory, the running time required to process a page is 0.087 second son average computed over a random selection of 100 pages.

We have also used page ranking method in order to bring out the expected data records so that the time consumption has been reduced. Accurate extraction of data records has been tested and the accurate results have been achieved. The performance of the data extraction methods is compared in three different ways. The other two evaluations focus on specific properties of the query result pages. Non contiguous QRR evaluation compares the performance for query result pages in which the QRRs are contiguous and noncontiguous.

V. CONCLUSION

Automatic data extraction, Efficiency in aligning the data which will be used in the application of Multiple Web data bases and the integration of the data can be achieved. It also only used for comparison shopping but also used in reducing the time consumption while extracting the data because the implementation of page ranking method has been introduced which speed up the search engine to extract the records and the clustering of the data regions are also implemented. Handling non-contiguous regions and nested structure processing difficulties are also rectified in this Updated CTVS method.

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 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, Jiyang wang, Frederick H. Lochovsky, "Combining Tag and Value Similarity for Data Extraction and Alignment" IEEE Transactions on Knowledge and Data Engineering, vol.24, No.7, July 2012.

[2] A. Arasu and H. Garcia-Molina, "Extracting Structured Data fromWeb Pages,"Proc. ACM SIGMOD Int'l Conf. Management of Data,pp. 337-348, 2003.

[3] 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.

[4] C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery,"Proc. 10th World Wide Web Conf.,pp. 681-688, 2001.

[5] 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.

[6] B. Liu and Y. Zhai, "NET - A System for Extracting Web Data fromFlat and Nested Data Records," Proc. Sixth Int'l Conf. WebInformation Systems Eng.,pp. 487-495, 2005.

[7] K. Simon and G. Lausen, "ViPER: Augmenting AutomaticInformation Extraction with Visual Perceptions,"Proc. 14th ACMInt'l Conf. Information and Knowledge Management, pp. 381-388,2005.

[8] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-AssistedData Extraction,"ACM Trans. Database Systems,vol. 34, no. 2,article 12, p. 35, 2009.

[9] C. Tao and D.W. Embley, "Automatic Hidden-Web TableInterpretation by Sibling Page Comparison," Proc. 26th Int'l Conf.Conceptual Modeling,pp. 566-581, 2007.

[10] Y. Zhai and B. Liu, "Structured Data Extraction from the WebBased on Partial Tree Alignment,"IEEE Trans. Knowledge and Data Eng.,vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[11] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "FullyAutomatic Wrapper Generation for Search Engines,"Proc. 14thWorld Wide Web Conf.,pp. 66-75, 2005.

[12] Duhan, N. "Page Ranking Algorithms: A Survey" Advance Computing Conference, 2009. IACC 2009. IEEE International.

