IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

Data Extraction and Alignment by using Combining Tag and Values Similarity

Aparna Pathak[1], Dr. Sadhana Chidrawar[2]

PG Student, Dept of CSE, MPGIN, Nanded, Maharashtra, India[1]

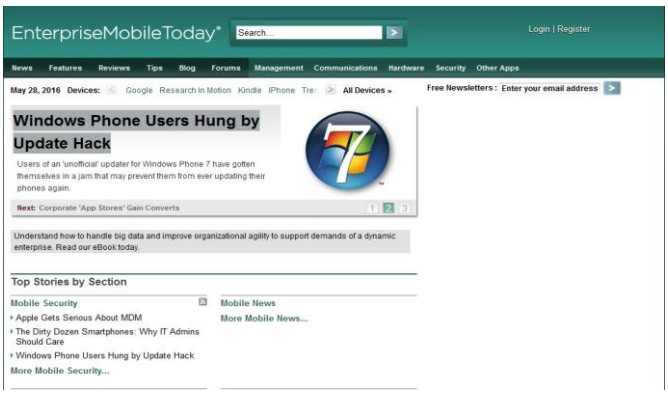
Assistant Professor, Dept of CSE, MPGIN, Nanded, Maharashtra, India[2]

Abstract: Based on the query entered by user web databases generate query result pages. Assistant Professor, Dept. of CSE, MPGIN, Nanded, Maharashtra, IndiaMany techniques are available to extract data and align it but they did not consider the unstructured data. This paper presents a novel technique for data extraction and alignment. It collects data from internet related to user's query which contains huge amount of data and it can be in unstructured form also. Here the aim is to find out only important data from it and align it in tabular form so that it will be very easy to compare different data.

Keywords: Data Records, Tag Tree Structure, Record Extraction and Records Arrangement.

I. INTRODUCTION

Online databases, called web databases, comprise the deep web. According to a query entered by user through the interface of a web database, pages in the deep web are generated dynamically, while web pages in the surface web can be accessed by a unique URL. A web database returns the relevant data, related to user's query which can be in any form i.e. structured, semi structured or unstructured, encoded in HTML pages. Many web applications, such as meta querying, data integration and comparative shopping requires the information from different web databases. Automatic data extraction is necessary for such type of applications to utilize the information and data hidden in different HTML pages. The collected data can be compared and aggregated easily only when it is extracted and organized in a structured manner, like tables. Hence, for these applications to perform correctly accurate data extraction is important. The problem of how to extract data records automatically from the query result pages generated by web databases is focused in this paper. 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. The main aim of data extraction from web pages is to remove any unnecessary information from the generated query result page, extract the query result records from the page, and align the extracted query result record into a tabular form. The main aim of proposed system is to drop the unwanted records from the query page, mine the QRR from the query page and arranged the mined QRR in a structured format. Query result record in web pages presents the host and pages imperative information which includes the list of product services, so it is important to mine these records.
11There have been lots of algorithms proposed on extracting the data automatically from web database in the past year. Fig. 1 shows the query result web page which is produced by web document.



II. LITERATURE REVIEW

Chang, Chia-Hui, Shao-Chen Lui [5], discovers the rules called IEPAD (Information Extraction Based on Pattern Discovery) for extracting automatic data from the web pages. This rule also identifies the records limited area by

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

mining the imitated pattern and numerous progression arrangements. It uses PAT (Patricia tree) tree for discovering the repeated pattern and it is further prolonged by arranging the pattern of all record occurrences.

ArvindArasu, Hector arcia-Molina [1],designed a new technique to the data extraction issue.It analyze many website uses the common template or layout for their respective web pages and extracts the database values from these template without learning any example or other similar human input.It designs an algorithm which takes the number of template generated web pages as a input,drop the unwanted information and extracts the data as output which is encoed in web pages.

Georg Lausen, Kai Simon [6], it describes the issue of unsupervised data extraction. Unsupervised data extraction becomes attainable when a query page contains the repetitive patterns. It generates the extraction rule by operating on DOM tree of the query page. It also explains the identification and rank probability of imitative patterns with respect to user's observable view point of the query page and it also match the sub-sequence of the pattern and they are arranged with global multiple series arrangement method.

ValterCrescenzi, Giansalvatore Mecca, Paolo Merialdo [2], it explains how to extract the data from HTML site, by using automatically generated wrappers. Wrappers is the program used for extraction of data from web pages. It develops the new technique that correlates HTML pages and produces a wrapper with respect to their similarities and variations.

Bing Liu, YanhongZhai [9], it explains the issue of automatic web data retrieval from several structured data records. They also explain how to segment the QRR, extracting the records from the data region and put them in Tabular form. They focus on two things those are, Data records recognition from the query page and next is arrange these extracted data in a table.

Robert Grossman, YanhongZhai, Bing Liu [4], mainly focused on the data record which contains the large amount of information on the web. Data records also contain the information regarding their host pages for example list of product or services. Mining the data records means extracting the information from them to provide value added services. It also explains mining of both contiguous and non contiguous data records.

In records recognition they segment the data record which helps to mine the data records from web page. In next step it uses the partial tree arrangement approach which is based on tree matching, it means arranging those data field in a couple of data records that can be arranged and these independent to other data field.

Jiying Wang, Fred H. Lochovsky [8], it describes the system which rebuilds the section of invisible back end database. It sends a query by using HTML form, and generates the regular expression wrappers that mine the data from query page and put the retrieved data in a structured format i.e. table.

III. SYSTEM OVERVIEW

CTVS is the new method which is specially designed to automatically retrieve the QRR from the query page. When user input any query to WWW, its web database generates the query result pages.

Our Proposed system is based on the two steps,

- Record Extraction
- Record Alignment

A. Record Extraction

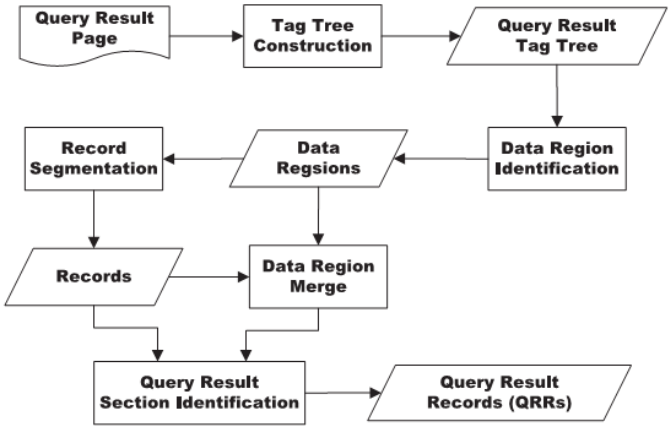


Fig. 1. Record Extraction Framework

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

In Record Extraction phase, firstly it identifies the data region which contains the number of query result records and then it does the segmentation of records [4]. Record alignment steps properly align the extracted data in a structured manner means it arrange the all the extracted QRR's in a table.

CTVS also useful for handling or extracting QRR's which is not contiguous, which may contain the other information such as comment, advertisement, recommendation etc. Our Proposed technique is mainly based on the tag structure to extract the data values.

Fig 2 shows the record extraction framework for the proposed system. Initially web database produces the query result web page based on the user given query, then <HTML> tag tree is build for the respective web page in Tag Tree building section. Data region recognition section recognizes the data regions that are present in query output page, data region consist of the record information that is to be retrieved. The purpose of record segmentation section is that it separates the recognized data region into data record based on the tag pattern style of the data region.

• Tag Tree Construction Phase

As we know that every webpage contains the HTML tags means web pages build by using the HTML tags. This module discovers all the data records which are formed by using <html>, <table>, , <form> tags and so on[3]. Before building tag tree for the respective query page, it becomes necessary to extract all the tags which are needed to build the query page. Tag tree is constructed from the root node that is HTML node to its child node which are used to design a web page.

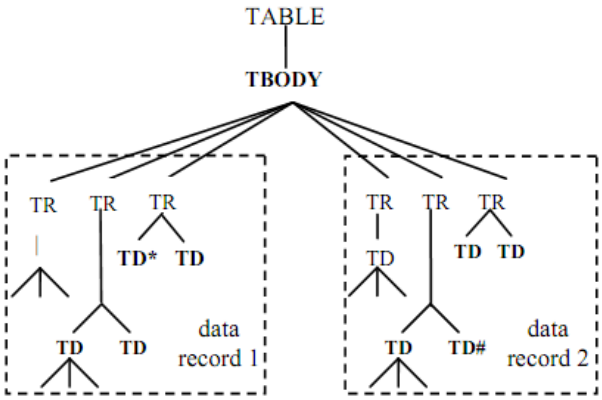


Fig. 2. Tag Tree of the page

• Data region identification

The tag tree building phase makes it easy to identify the data regions from that query page. In this section first we recognize the data regions in query page which contains numerous data records [4][7]. Some child sub tree of the same parent node, here node is nothing but the HTML tags which forms data regions which is having data records. We find the data region by combining and comparing multiple adjacent tag string of individual nodes. Mining is done in a page that contains similar data record.

• Record Segmentation

In record segmentation module, it finds the tandem repeats (repeated tag substrings) that are present in data region or not. In data region if any tandem imitate is raised, we consider that each imitated occurrences inside the tandem repeat related to a record, if many tandem imitates are there in a data region, then we required to choose one to imply the record.

B. Record Alignment

After extraction of record, we align the extracted records in a structured form. This phase arranges the records in structured form by analyzing the trees for every mined query record. We align the record by using three methods,

• Pairwise Record Alignment

In pair wise record arrangement, the data field are arranged in combination of data query result record which presents the way for how the data field should be arranged among all the records. The proposed system is depends on the perception that the data items corresponding to equivalent field, they must have type of data and also contains the same string, since the query result records for the similar query.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

• Global Data Record Alignment

Global data records alignment, arrange the QRR in a table globally, in this all the data items of the similar attributes are arranged in a same field of table. In this alignment we denote vertex as a data values of QRRs and pair wise arrangement is treated as edges, in which pairwise arrangement method can be illustrated as undirected graph. The problem in global alignment is that finding the connected component in a undirected graph. In undirected graph each joined component represents the table field in which all the joined data items from other QRRs are then arranged vertically.

• Nested Structure Processing

The purpose of this method is that it recognizes the data item of a query result record that is discovered by this method. The case when QRR contains the multi valued attribute, then a few of the data values may not be arranged to other data values. The proposed system does not use the this type of arrangement before the data records are arranged as it is aligned in DeLa [8] and NET [9], it uses it later the data records are arranged. Using this arrangement before the data records are arranged, it makes them unsafe to optional attribute so due to what it makes the tag structure irregular. Our proposed system avoids this problem. The data value similarity information effectively prevents flat structure from being identified as a nested structure in CTVS. As it shares similar tag structures, a structure with several columns having the same tag structure, may be identified as a nested structure mistakenly in DeLa. Such wrongly identified flat structure can have serious effects. DeLa groups all the values into one parent and then aligns them to other records, making the arrangement much more difficult.

IV. EXPERIMENTAL RESULTS

A. Dataset

In this paper, we have worked on the dataset called TBDW version 1.02 [11], which is having 51 online databases. For this five query pages are formed for each web database.

B. Performance Measures

We use two familiar dimensions and recall, to check out the execution of this system. Record level recall is the ratio of number of QRR that have been appropriately mined over the total number of QRR in a query page. Record level precision is the ratio of the number of record that have been appropriately mined over the total number of records that have been mined.

C.Result and Analysis

Table I - Comparison between CTVS and DeLa

Parameter/Method	CTVS	DeLa
Extracted Records	680	643
Correctly Extracted Records	673	594
Record Level Precision	93%	88%
Record Level Recall	94%	84%
Page Level Precision	90%	73%

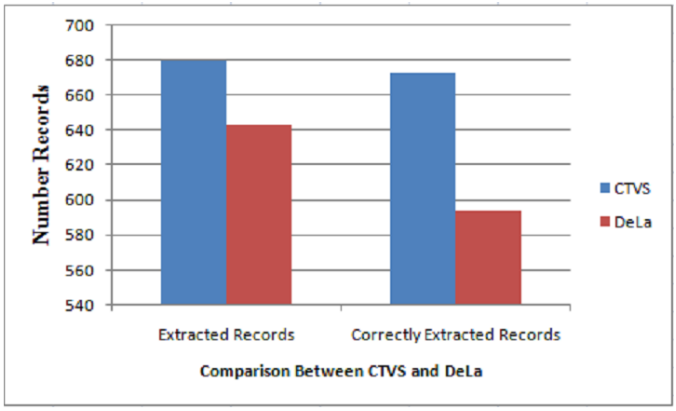


Fig. 3. Record Extraction comparison graph for CTVS and DeLa

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2017

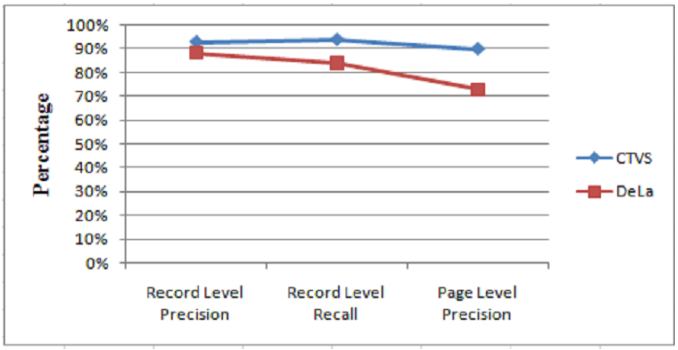


Fig5 Comparison Graph between Our Approach and DeLa for Record and Page Level Recall and Precision

Figure 4 and Figure 5 shows the comparison results between our approach and DeLa system. Here we present the experimental result for CTVS over the TBDW dataset and also we compared results with DeLa system [8]. By comparing CTVS with DeLa system, we come to know that CTVS consistently has best performance metrics over TBDW dataset.

V. CONCLUSION

Combining tag and value similarity (CTVS) technique is specially used for knowledge discovery purpose. CTVS method works on two principles. The first principle is that, it recognizes the data region available in the query page and after that segments it. Our proposed system also works on the non contiguous data which is present in the data regions. The second principle is that alignment of data values in a structured manner. The alignment of data value is done by using three methods: First one is by pairwise, second one is by holistic and last one is by nested structure processing. CTVS system uses the tag structure to find the data values. If query page contain numerous data regions which is having data records and these data records checked with records which is in another data region if these records are not similar, then CTVS choose only one data region and discards others.

REFERENCES

[1] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.
[2] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 109-118, 2001.
[3] Y. Zhai and B. Liu, "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.
[4] B. Liu, R. Grossman, and Y. Zhai, "Mining Data Records in Web Pages," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 601-606, 2003.
[5] C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proc. 10th World Wide Web Conf., pp. 681- 688, 2001.
[6] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions," Proc. 14th ACM Int'l Conf. Information and Knowledge Management, pp. 381-388 2005.
[7] 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.
[8] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th World Wide Web Conf., pp. 187-196, 2003.
[9] B. Liu and Y. Zhai, "NET - A System for Extracting Web Data from Flat and Nested Data Records," Proc. Sixth Int'l Conf. Web Information Systems Eng., pp. 487-495, 2005.
[10] Weifeng Su, Jiying 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,pp. 1186- 1200, July 2012.
[11] http://daisen.cc.kyushu-u.ac.jp/TBDW/.

BIOGRAPHIES

Aparna Pathak, M.E.CSE, MPGI Nanded.

Dr. Mrs. Sadhana Chidrawar, Professor in CSE dept, MPGI Nanded,